Development of a Deep Learning-based Tobacco Plant Counting Algorithm through Aerial Imagery using Object Detection and Segmentation Techniques

Author

Ramsha Shahid

00000320725

Supervised by

Prof. Dr. Umar Shahbaz Khan

MASTERS IN MECHATRONICS ENGINEERING,

DEPARTMENT OF MECHATRONICS ENGINEERING,
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING,
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD, PAKISTAN.

JULY 2023

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis by Ms. **Ramsha Shahid** Registration No. **00000320725**, of Electrical and Mechanical Engineering College has been vetted by undersigned, found complete in all respects as per NUST Statues/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature:

Name of Supervisor: Dr. Umar Shahbaz Khan

Dated: 19/7/23

Signature of HOD:

Dr. Amir Hamza

Date: 19/7/23

Signature of Dean: Rashid

Brig Dr. Nasir Rashid

Date: 19 JUL 2023

Dedicated to my exceptional parents, adore siblings, and friend Zahra their endless support and encouragement led me to this wonderful accomplishment

Acknowledgments

I am grateful to Allah Who is the most Merciful for giving me the ability to accomplish my thesis.

I am grateful to my cherished guardians and my adored siblings for their help to me in every period of my life.

I might likewise want to communicate unique gratitude to my supervisor Dr. Umar Shahbaz Khan and Dr. Waqar Shahid for their assistance throughout the duration of my thesis. I can confidently affirm that without their assistance, the completion of this research would not have been possible.

I would also like to extend my heartfelt gratitude to my classmates Sidra Muqaddus and Fahad ul Hassan Asif Mattoo for their endless support and motivation. Their willingness to share their knowledge fostered a supportive learning environment.

I would like to express my sincerest gratitude to Dr. Syed Imran Moazzam and NCRA for their invaluable collaboration in data collection and sharing.

I am also thankful to Ayaz Fayyaz and all staff of the Mechatronics Engineering department for their support and cooperation.

Abstract

Accurately estimating crop emergence prior to harvesting is increasingly critical for ensuring the long-term sustainability of natural resources. This process serves multiple purposes, including yield estimation, seed quality prediction, identification of regions prone to yield losses, and formulation of effective agricultural plans. By maximizing crop population within the constraints of limited land and resources, crop emergence estimation contributes to the sustainable utilization of these valuable resources. However, existing plant counting frameworks often require extensive offline image processing using licensed software to generate orthomosaics through the multiview stereo, resulting in significant computational demands. To address this challenge, this study proposes a comprehensive plant counting framework that directly estimates plant counts from aerial images. The framework comprises three essential modules: overlap detection, plant detection, and plant counting. The overlap detection module eliminates the need for computationally intensive orthomosaic generation by utilizing only visual cues to mask overlapping areas, thereby preventing duplicate plant counting. Three distinct methods are evaluated as core modules to identify an optimal generalized solution for plant counting, considering both time complexity and accuracy. The first method employs semantic segmentation with U-NET after overlap detection for plant detection followed by counting connected pixels. In the second method, object detection using YOLOv7 is utilized for plant detection after overlap removal. Finally, the third method introduces a real-time plant counting framework based on multiple object tracking, employing YOLOv7 for object detection and SORT for object tracking as a replacement for the overlap detection module. The proposed algorithm is evaluated using high-resolution aerial data collected from two separate Tobacco fields near Peshawar, Pakistan. The first and second methods achieve average F1 scores of 0.947 and 0.9667, respectively. Notably, the third method exhibits promising potential for real-time applicability, achieving an average F1 score of 0.967.

Keywords: *Semantic Segmentation, Plant Count, Deep Learning, U-Net, Overlap Detection, Object Tracking, Object Detection, YOLO, SORT*

Table of Content

List of Figures

List of Tables

List Abbreviations

DL	Deep Learning
ML	Machine Learning
SVM	Support Vector Machine
YOLO	You Look Only Once
SORT	Simple Online and Real-time Tracking
IOU	Intersection over Union
UAV	Unnamed Aerial Vehicle
MAP	Mean Average Precision

Chapter 1: Introduction

1.1 Overview

Agricultural countries highly depend upon their agriculture for economic growth. However, this sector is also dependent on natural resources like land and water, which are becoming increasingly scarce due to the growing demands of a rapidly increasing population. It's therefore imperative to sustain these resources by optimizing their usage.

One of the ways to optimize the usage of natural resources is by estimating crop emergence during the initial stages of growth. Crop emergence refers to the emergence of a plant from its seed, and it's a critical stage in crop production. At this stage, every seed that germinates has the potential to become a productive plant, leading to maximum yields. As we can apply agricultural input like reseeding etc. to identify potential yield losses at this initial stage. Accurate estimation of crop emergence is important because it allows farmers to assess the success of their planting, make timely management decisions, and forecast crop yields. On the other hand, if the seeds fail to germinate or emerge weakly, it can result in lower yields, reduced productivity, and wasted resources.

This research presents a generalized framework for plant counting. The proposed generalized framework consists of Deep Learning modules for identification of Plants and a novel overlap detection module. We have evaluated multiple DL models to find the best for this framework. Other than this we also evaluated Object Detection with Tracking for real time plant counting. The results are evaluated for Tobacco Fields.

1.2 Importance of Plant Counting in Pakistan

In Pakistan, agriculture is the backbone of the economy, contributing around 20% of the GDP and employing over 40% of the labor force. The country is known for its production of cotton, wheat, rice, sugarcane, and maize, among other crops. However, agriculture in Pakistan is vulnerable to various challenges, including water scarcity, soil degradation, pests and diseases, and unpredictable weather patterns.

Crop emergence estimation can help farmers in Pakistan to address some of these challenges. For example, if emergence rates are low, farmers can take corrective measures such as reseeding or adjusting irrigation to ensure that crops reach their full potential. Early identification of emerging problems can also help farmers to reduce the risk of crop failure, resulting in increased crop yields and improved food security.

Moreover, accurate crop emergence estimation can also assist in making informed decisions regarding the use of inputs such as fertilizers, pesticides, and herbicides. By knowing the percentage of crops that have emerged, farmers can apply these inputs more efficiently and reduce their costs while minimizing the impact on the environment.

1.3 Challenges in Plant Counting

Traditionally plant counting is done manually through human labor that is error prone, time consuming and labor-intensive task. Manual plant counting can result in poor identification of areas having potential yield losses. This leads to ineffectiveness in the process of crop emergence estimation. Cutting edge technologies Deep Learning (DL), Computer Vision etc. utilization for plant counting is proven to be a reliable solution. These applications require the data of the entire field. This large-scale field monitoring can be possible with unmanned aerial vehicles (UAV) imagery or through sentinel data. Sentinel data unavailability, polygon formation, cloud calibration and storage capacity make it ineffective as compared to UAV imagery that can be acquired at any favorable time and weather conditions making it a feasible solution for real time plant counting. UAV imagery at any framerate results in the overlap between frames that results in inaccurate redundant plant count. So, state of the art techniques solved this overlapping problem through Orthomosiac formation, discussed as follows:

1.3.1 Orthomosiac Formation

An orthomosaic is a high-resolution, georeferenced, and orthorectified image created by stitching together multiple overlapping aerial photographs or images. The resulting orthomosaic provides a detailed and accurate representation of the area captured in the images, with distortions and perspective effects removed.

Orthorectification is the process of removing the effects of terrain relief and camera perspective distortions from the original images, such that the resulting mosaic can be used for accurate measurements and mapping applications. To achieve this, a digital surface model (DSM) of the terrain is generated from the original images, which provides information about the height of the objects captured in the images. The DSM is used to correct for variations in elevation, such as hills and valleys, that can cause distortions in the images. The process involves warping and blending the images to create a seamless and georeferenced orthomosaic making it a computational expensive process. There are many software packages available to perform these steps, including Pix4D, Agisoft Metashape, and DroneDeploy, among others. However, creating a high-quality orthomosaic requires expertise in remote sensing, photogrammetry, and GIS.

1.4 Tobacco Fields in Pakistan

There are more than 50,000 tobacco growers in Pakistan. Every year many fields in hectares are occupied by Tobacco fields. It is of great importance that every bit of land occupied, and resources used must be economically at most beneficial.

Table 1.1: Tobacco fields in Hectares

Year	**Flue-cured Virginia**	**Dark air-cured**	**Rustica**	**Whaite Patta**	**Burley**	**Total**
2013-14	27413	1225	16004	4341	57	49040
2014-15	30765	925	16822	5250	42	53804
2015-16	29061	872	17434	5278	40	52685
2016-17	26121	599	16609	3880	38	47247
2017-18	24527	1367	19025	1366	47	46332
2018-19	24790	740	17702	630	56	44877
2019-20	27639	896	21201	1003	50	50789
2020-21	27150	591.6	10959	1190	50	39941
2021-22	23159	586	9823.62	1106	50 +66 (Sun-cured Virginia)	34790.62

1.5 Research Objectives

The research objectives to achieve the primary goal of creating a framework for effectively estimating crop counts in various agricultural scenarios are as follows:

- Develop a generic plant counting framework based on Computer Vision with Deep Learning (DL) that is applicable to multiple crop stand counts.
- Explore and implement innovative algorithms/models within the framework.
- Replace the computationally expensive orthomosaic formation by implementing overlap detection based solely on visual cues.
- Eliminate the dependency on commercialized/licensed products for orthomosaic formation.
- Evaluate the effectiveness of multiple object tracking on UAV imagery for plant counting in order to develop a real-time system

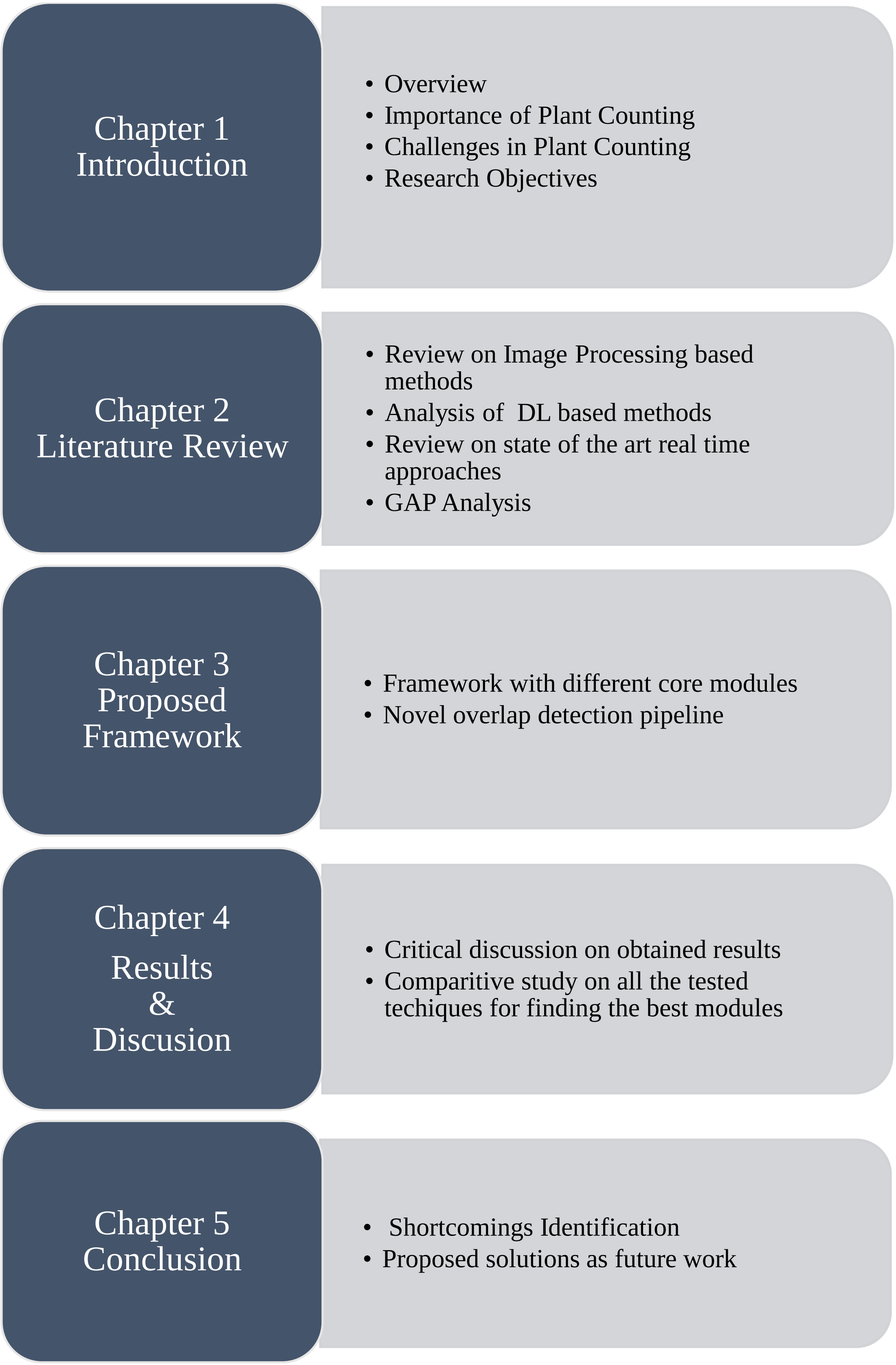

Chapter 1
Introduction
Overview
Importance of Plant Counting
Challenges in Plant Counting
Research Objectives
Chapter 2
Literature Review
Review on Image Processing based methods
Analysis of DL based methods
Review on state of the art real time approaches
GAP Analysis
Chapter 3
Proposed Framework
Framework with different core modules
Novel overlap detection pipeline
Chapter 4
Results & Discusion
Critical discussion on obtained results
Comparitive study on all the tested techiques for finding the best modules
Chapter 5
Conclusion
Shortcomings Identification
Proposed solutions as future work

Figure 1.1: Flowchart of the Research

Chapter2: Literature Review

In literature, precision agriculture being a hot domain in research, diverse techniques are found for one its sub domain that is plant counting. These diverse techniques involve Image Processing, Computer Vision, Machine Learning (ML) and Deep Learning for plant identification followed by counting.

2.1 Image Processing based Plant Counting

Image processing-based plant counting involves contour detection, segmentation, thresholding based on vegetation indices.

2.1.2 Contour Detection

Contours are the outlines of the objects that can be represented as a sequence of connected points or curve. Contour detection is the detection of boundaries around an object. Contour detection involves techniques like thresholding, edge detection and segmentation. A method based on image processing is proposed [1], which involves detecting contours followed by combination of morphological operations.

2.1.2 Watershed Segmentation

Watershed segmentation is a popular image segmentation technique that separates an image into distinct areas based on the gradient magnitude topology. The main principle underlying watershed segmentation is to treat the image as a topographic map, with each pixel's brightness corresponding to the terrain's elevation. Plant region extraction using watershed segmentation is proposed in [2], extracted plant region consist of Tobacco and non- Tobacco regions. These are classified using CNN architecture followed by post processing. The proposed algorithm consists of several pre-processing and post-processing steps.

2.2 Segmentation Through Vegetation Indices

Vegetation indices are mathematical formulas used to analyze remote sensing data, particularly from satellite imagery, to estimate the density, health, and productivity of vegetation. These indices are based on the principle that plants absorb and reflect different wavelengths of light in varying amounts, and by measuring the reflectance of certain wavelengths, we can infer certain characteristics of the vegetation. A segmentation based on vegetation indices was proposed in [3] , this study utilized EXG (Excess Green Index), NGRDI (Normalized Green-red Difference Vegetation Index), and (EXG-EXR Excess Green Minus Excess Red Index) for detection of Tobacco Plants in an Orthophoto. After segmentation a high pass filter is utilized to suppress noise. The technique showed promising results but requires high computational power. Combination of EXG and Otsu Threshold is proposed [4] for the extraction of wheat seedlings. The resulting images contain holes and noise that is resolved using morphological operations. Six Vegetation Indexes were utilized by [5], they have utilized Random Forest for the prediction of soybean plant. The model was trained on 66 experimental plots collected in 2018 of soybean whereas test dataset contains 200 plots from the year 2019. In the proposed approach [6], the extraction of features for wheat ears segmentation and feature extraction from an orthophoto was accomplished using Laplacian and Finding Maxima filters. Once the features were extracted, a machine learning-based approach was employed for classification. Among the tested algorithms, Support Vector Machines (SVM) and Random Forest (RF) demonstrate superior performance compared to other machine learning algorithms. These algorithms effectively learn and classify the extracted features, accurately distinguishing between wheat ears and background elements.

2.3 Plant Counting Using Deep Learning

Deep learning being cutting edge technology have been utilized in almost every sector. The current state of the research involves mainly object detection and semantic segmentation.

2.3.1 Object Detection

Object detection framework was utilized in [6] on very high-resolution imagery. The Structure from motion (SFM) was generated using licensed software in order to get camera internal orientation an external orientation. Then, You Only Look Once (YOLOv3) was trained to detect the plants in original image. The image was cropped before detection in order to detect the smaller seedlings. Lastly, the image coordinates were projected to geographic coordinates. A deep learning-based method for cotton boll counting is proposed using multi-receptive filed extraction called MRF-YOLO[7]. It consists of a multi-scale residual block and attention module to enhance the feature details and loss is reduced using a multi-receptive field extraction module followed by a small target detection layer for the improvement detection of precision. A CNN for wheat head detection called WheatNet [8] was proposed for wheat head counting using point annotation. Truncated MobileNetV2 was used as a feature extractor followed by a two branched architecture for localization and counting. Experimental evaluations demonstrate the effectiveness and superior performance of WheatNet compared to other existing methods for wheat head detection

2.3.2 Segmentation

A pipelined framework was proposed in [9], that consist of distortion removal, crop row detection, geo referencing rows and followed by pre-trained RESNET18 architecture for crop emergence and canopy size estimation. The framework was evaluated for per image estimation. A system was developed [10] which employs deep neural networks and geometric descriptors to estimate the number of early-season maize plants in low spatial resolution aerial data. The raw data was combined to form a single large image, and the Max Area Mask Scoring RCNN was used to detect each row. The detected row was then horizontally rotated, followed by segmentation of soil and green plants. The sparse region detection algorithm was applied to each row, and the results were combined for all detected rows to calculate the average plant stand count of the entire field. The study[11] focused on the semantic segmentation of sorghum using hyperspectral data and identification of genetic associations. The researchers aimed to develop a method that could accurately segment and classify different components of sorghum plants based on their genetic characteristics. Organ-level semantic segmentation presents promising opportunities for identifying genes that influence variation in various morphological phenotypes across grain crops like sorghum, maize, and related species. Valente et al.[12] employed the Otsu thresholding method for segmenting spinach plants. They first converted the orthomosaic into smaller units or patches and then applied the Otsu thresholding algorithm to distinguish the plant pixels from the background. By effectively thresholding the image, they were able to separate the spinach plants and extract them for further analysis. To assess the number of pixels per plant, utilized the AlexNet architecture, a deep convolutional neural network (CNN) model known for its excellent performance in image classification tasks. By leveraging the capabilities of AlexNet, they were able to analyze the segmented spinach plants and estimate the number of pixels dedicated to each individual plant. A two-branched CNN-based architecture [13] was evaluated for per-image plant counting in orchard and corn fields. This architecture comprised two separate branches, each dedicated to different aspects of plant counting. The first branch focused on plant localization, accurately identifying the positions of individual plants within the image. The second branch was responsible for counting the detected plants per image. RiceNet [14]for detection and localization of rice plants was proposed that consist of multiscale feature fusion with plant attention mechanism that outer perform on the URC dataset.

By accurately segmenting and classifying different organs or plant structures within these crops, such as leaves, stems, panicles, and other components, researchers can study the genetic associations underlying diverse traits. Nee et al. [15] utilized U-Net for semantic segmentation on UAV data after row detection using Hough Transform for row detection. The identified rows than followed by semantic segmentation to segment corn plants.[16] also utilized U-Net for semantic segmentation of very high-resolution imagery from different altitudes. U-Net results are then followed by morphological operations and blob detections for plant counting.

2.3.3 Object Detection and Tracking

Multiple object tracking was utilized [17]for cotton emergence estimation. Modified CenterNet was used for object detection whereas DeepSORT[18]. This proposed technique was evaluated on aerial images taken from a height of 0.5m.

YOLOv4 in combination with DeepSORT is utilized for the development of Multi-object framework[19] for pear fruit detection and counting. The proposed framework was evaluated using different versions of YOLOv4. YOLOV4 for detection and optical flow for tracking for cotton seedling is utilized in the study[20].The integration of YOLOV5 and DeepSORT is applied [21] to detect and localize the generative organs, such green tomato, red tomato and flowers in the images. Tomato data was acquired in an experimental setup and resulting F1 scores of red tomato, green tomato, and flower classes are 0.74, 0.56, and 0.61, respectively. UAV system is proposed [22]to detect, localize and count ornamental plants in their natural habitats.

Although object detection and semantic segmentation in DL-based approaches show promising results, they involve computationally expensive processes such as orthomosaic formation or row detection based on GPS location. Other than skipping the orthomosaic formations step based multi object tracking frameworks for counting one study [9] developed a framework for real-time cotton stand count and canopy size mapping. The cotton rows in each image frame were detected and the row angle was used for rotating each individual frame. The row spacing was used as a reference for dynamic GSD calibration of each image frame. Seedlings in every individual frame were located based on their position in image coordinates, the GSD and through georeferencing cotton rows to replace orthomosaic formation.

2.4 Research Gap Analysis

The conventional approach of orthomosaic formation in cutting-edge research typically relies on resource-intensive computational power and commercial softwares such as Agisoft PhotoScan, Zephyr, and Pix4D. This offline processing method limits the advantages of real-time plant counting. Recognizing the promising outcomes of utilizing UAV imagery in advanced agricultural fields, this study endeavors to develop a versatile plant counting framework.

The proposed framework introduces a novel overlap detection module that solely relies on visual cues, eliminating the need for computationally demanding orthomosaic formation. By combining the overlap detection module with semantic segmentation and object detection techniques, the feasibility of estimating tobacco plant counts is evaluated. The results obtained through this framework demonstrate significant time savings, thereby enhancing the benefits of plant counting. These benefits include the optimized allocation of resources during the early stages of crop growth, enabling farmers to make informed decisions regarding irrigation, fertilization, and pest management.

Moreover, the integration of an object detection algorithm with SORT tracking offers a real-time applicable method for plant counting. A real-time plant counting framework holds the potential to facilitate the early detection of plant diseases, pests, and other issues, thus mitigating the risk of crop losses. By promptly identifying such issues, farmers can implement appropriate preventive measures to ensure the health and productivity of their crops.

Chapter 3: Proposed Framework

A generalized framework that can replace the computational expensive process of orthomosaic formation is proposed as shown in Figure 3.1. The framework consists of a detection module for which object detection and semantic segmentation is evaluated. Another real time system is also proposed based on object detection in conjunction with multiple object tracking.

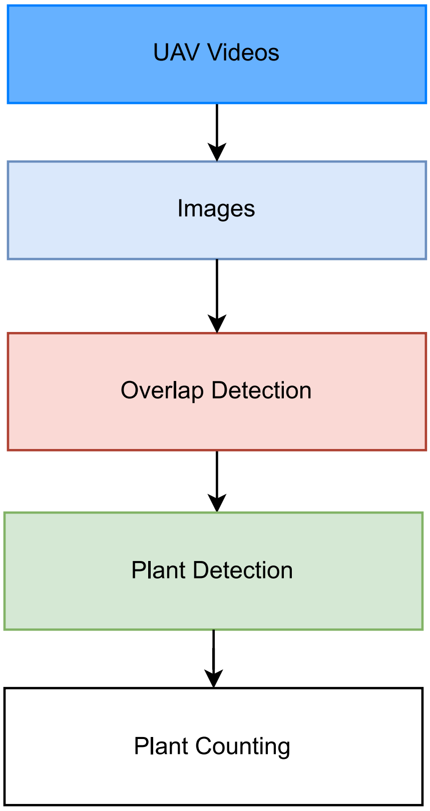


Figure 3.1: Proposed General Framework

3.1 Data Collection

We have obtained a new dataset of tobacco fields in Peshawar, Pakistan using a DJI Mavic Mini drone equipped with a high-resolution RGB camera, recording aerial data at a rate of 20-30 frames per second. The dataset comprises two fields of tobacco plants captured during the early growth stage, approximately 15-40 days after planting, at a resolution of 1920 x 1080 pixels. The two distinct tobacco fields are recorded at a rate of 20-30 frames per second. This resulted in the creation of two separate field videos, each representing a different tobacco field. To facilitate evaluation and ground truth analysis, these videos are further divided into smaller, distinct clips. The clips are given specific names, namely DATASET-1, DATASET-2, DATASET-3, DATASET-4, and DATASET-5, which is consistently used throughout the study to refer to the respective datasets. This naming convention ensured clarity and consistency when referring to specific clips during analysis and discussions.

Each dataset captured the movement of the drone in a single direction, maintaining a consistent flight path throughout the recording. This one-directional movement allowed for easier analysis and comparisons between datasets, as the drone's motion was consistent and predictable within each clip. Images of the tobacco dataset is shown in Figure 3.2, depicting different soil textures and sunlight conditions. This stage is considered optimal for plant counting as the correct agricultural inputs can boost production, and weed infestation is relatively low. However, due to manual control of the UAV, there may be speed and height variations, resulting in uneven frame overlap. To avoid counting repetition caused by overlap, an overlap detection technique has been devised.

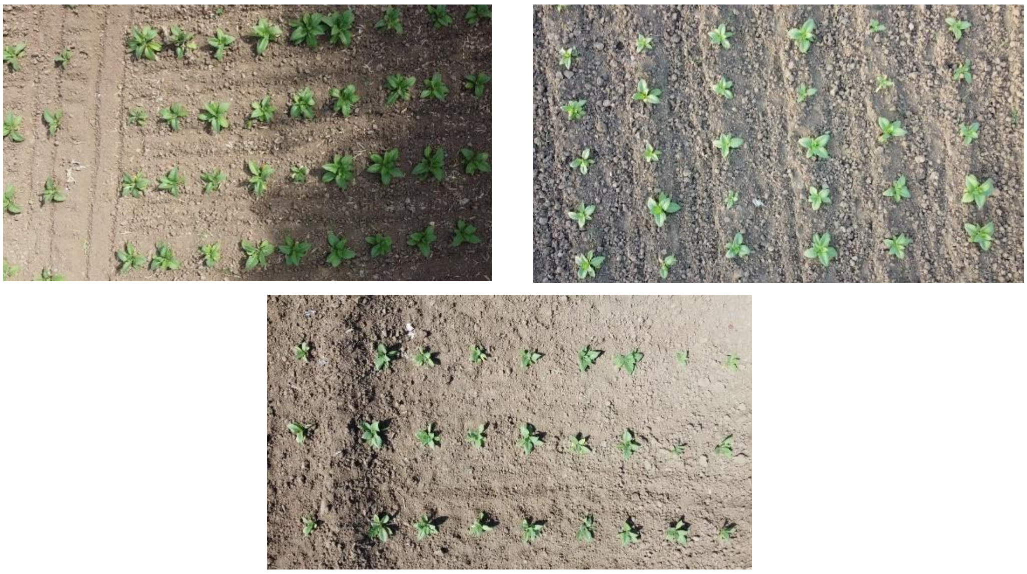


Figure 3.2: Samples of Acquired data set under different sunlight and soil conditions.

3.2 Overlap Detection

In this proposed framework, frames are extracted from video clips with a 40-50% overlap between consecutive frames. The detection of overlapping regions between frames is crucial for accurate plant counting. To ensure reliable counting results, it is necessary to accurately mark and determine the overlapping regions. By detecting and delineating these regions, potential issues such as double-counting or miscounting of plants can be avoided.

The process begins with feature extraction, where distinctive features are extracted from the frames. These features act as descriptors for identifying corresponding regions across frames. Next, feature matching is performed to establish correspondences between features from adjacent frames. Once correspondences are established, perspective transformation is applied to align the overlapping regions correctly. This transformation compensates for any perspective distortions or differences in viewpoints, ensuring accurate registration of the overlapping areas. This enhances the accuracy of plant counting by addressing the challenges posed by overlapping regions.

3.2.1 Feature Extraction

Feature detection and description is an active area of research in computer vision. It involves obtaining features that are highly distinctive and repeatable against various image transformations, which is crucial in many applications. The two most popular algorithms for multiscale feature detection and description are SIFT[23] and SURF. SIFT features use a Difference of Gaussians operator applied through a Gaussian scale space to obtain feature locations and build a descriptor vector of 128 elements based on gradient orientation. SURF features are inspired by SIFT and can be computed much faster using the integral image. They use a rectangular grid of 4x4 subregions and a sum of Haar wavelet responses weighted by a Gaussian centered at the interest keypoint to build a descriptor vector of 64 or 128 elements. The Gaussian scale space and sets of Gaussian derivatives are commonly used for scale space analysis in both the approaches and their related algorithms. However, it should be noted that Gaussian scale space does not preserve the natural boundaries of objects and equally smooths details and noise at all scale levels.

We have employed KAZE features[24], a technique for detecting and describing multiscale 2D features in nonlinear scale spaces. Unlike previous methods that depend on the Gaussian scale space, our approach is based on nonlinear scale spaces utilizing efficient Additive Operator Splitting (AOS) techniques and variable conductance diffusion. Although our method incurs a slightly higher computational cost, our results demonstrate significant improvements in both detection and description performance compared to previous state-of-the-art methods.

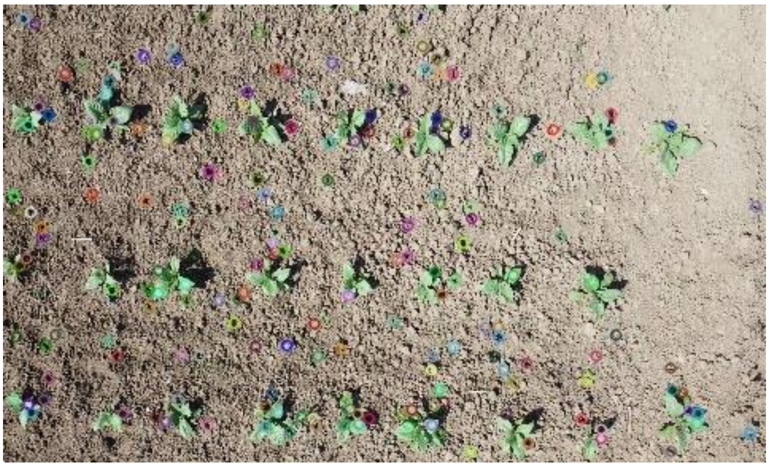


Figure 3.3: Feature Extraction with KAZE

3.2.2 Feature Matching

Feature matching is the process of comparison of features across images. We get a set of pairs of matching features for two images. As discussed in previous section we have extracted KAZE feature. So, to match the KAZE features of two successive frames, a (Fast Library for Approximate Nearest Neighbors) Flann-based matcher is employed. This matcher comprises a collection of algorithms that are specifically designed for speedy nearest-neighbor searches in high-dimensional features and large datasets. It is faster than (Brute Force) BF-based Matcher for significant datasets. The matches obtained from this process are sorted based on Euclidean distance, and the top 500 matches are deemed as good matches. These good matches are then utilized to estimate homography. Not all of the matched keypoints may be relevant. To separate the inliers from the outliers, the RANSAC (Random sample consensus) algorithm is used. The feature matching with two consecutive frames is shown in Figure 3.4.

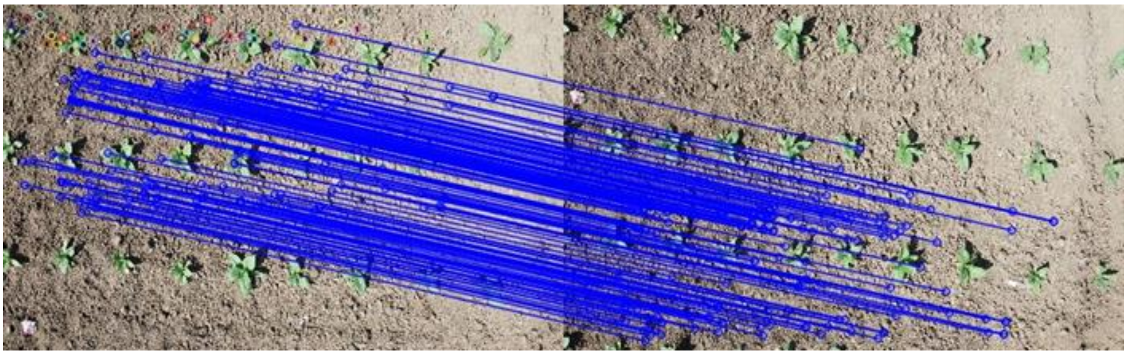


Figure 3.4: Feature Matching

3.2.3 RANSAC

RANSAC (Random Sample Consensus) [25] is a robust method used to estimate models in the presence of noise and outliers. It provides a reliable approach for handling datasets that contain erroneous or inconsistent data points. The RANSAC procedure can be divided into three key steps, each contributing to the robust estimation process.

The first step involves sampling the dataset into smaller subsets, treating these samples as inliers. This random sampling strategy helps in identifying potential data points that conform to the underlying model. By considering these samples as inliers, RANSAC reduces the influence of outliers in the estimation process.

In the second step, the model is estimated using the selected inliers. The algorithm computes a model based on the subset of inlier points, aiming to find the best-fit representation of the underlying structure within the data. The estimated model captures the relationship between the data points, despite the presence of noise and outliers.

In the third step, the algorithm calculates the score of inliers and outliers for the estimated model. Each data point is evaluated based on its fit to the estimated model, allowing for the differentiation between inliers (data points consistent with the model) and outliers (data points deviating from the model). This scoring mechanism helps to further refine the estimation by emphasizing the contribution of reliable inliers and reducing the impact of noisy outliers.

These three steps are repeated iteratively, typically over multiple iterations, to identify the model with the highest number of inliers. The number of iterations required depends on factors such as the probability of inliers, the probability of outliers, and the minimum number of samples necessary for accurately estimating the model.

3.2.4 Homography

The source points are determined by selecting the keypoints of sorted matches from the first frame or image, while the destination points are identified by selecting the keypoints of sorted matches from the second frame or image. These points are then used to compute the Homography, which relates the images of a plane captured by different camera orientations or positions Figure 3.5 shows the homography between consecutive frames.

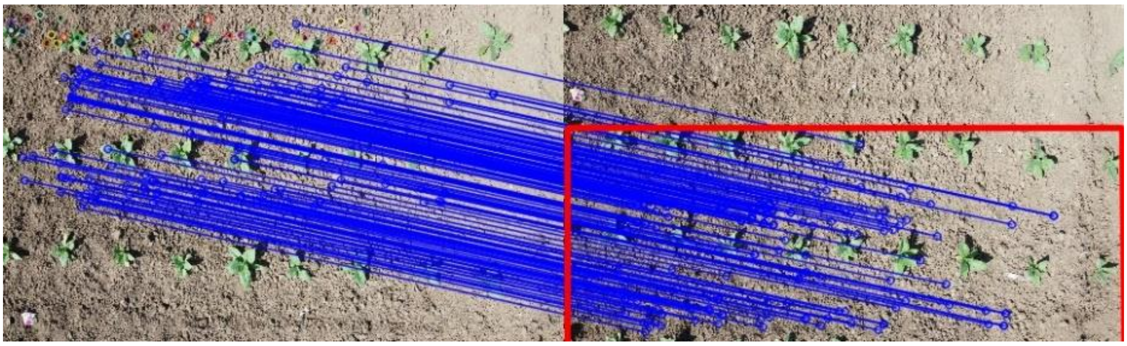


Figure 3.5: Overlap region identification with homography

Homography is represented by a 3-by-3 matrix (H) in homogeneous coordinates and can be calculated for each match using RANSAC to obtain the solution with the least number of outliers.

3.2.5 Overlap Removal

The process of identifying the overlapping regions between consecutive frames using the homography matrix, which is obtained through the previously described steps is shown in Figure 3.6. Once the overlapping region is identified, it is masked. This process effectively identifies and removes the overlapping areas from the image.

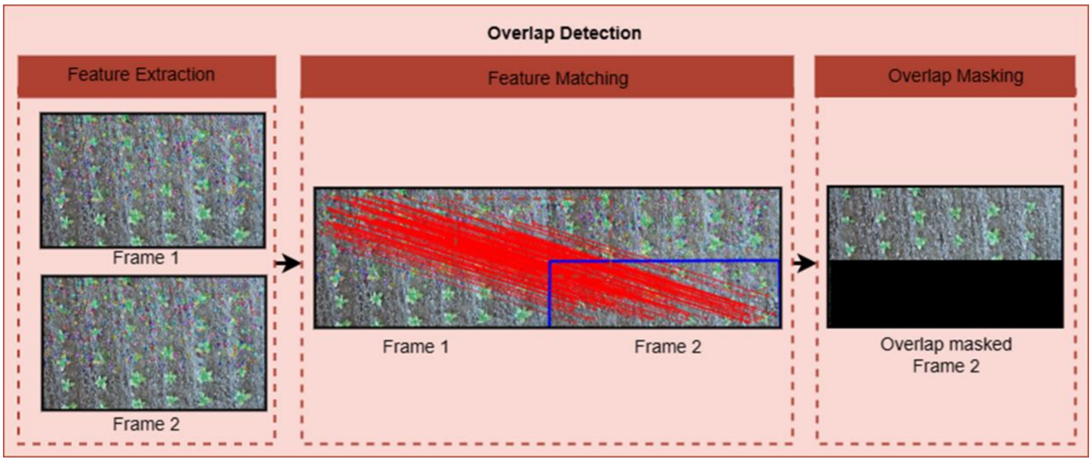


Figure 3.6: Overlap detection and masking pipeline

3.3 SVM as a Detection module

To find the optimal solution for plant detection for the proposed general framework we have utilized multiple models including the fundamental classification model Support Vector Machine (SVM) that lies in the realm of Machine learning(ML).

3.3.1 SVM

Support Vector Machine (SVM)[25] is a fundamental algorithm that holds great importance for machine learning experts. One of the reasons for its popularity is its ability to achieve remarkable accuracy while requiring relatively less computational power. SVM is a versatile algorithm that can be applied to both regression and classification tasks, although it is predominantly used for classification objectives. Its effectiveness and reliability in solving classification problems have made it a preferred choice in various domains.

SVM operates by constructing hyperplanes in a high-dimensional feature space to effectively separate different classes of data points. It aims to find an optimal decision boundary that maximizes the margin between classes, leading to improved generalization capabilities and better classification performance on unseen data. By utilizing a subset of training data points called support vectors, SVM can efficiently classify new instances based on their proximity to the decision boundary.

3.3.2 Proposed Framework with SVM

In the proposed general framework, we conducted an evaluation of Support Vector Machine (SVM) as the plant detection module, as depicted in Figure 3.7. The SVM module was trained on descriptors extracted from KAZE features, which are robust and distinctive image features commonly used in computer vision tasks.

To train the SVM, we manually labeled 4000 descriptors from the dataset. Each descriptor is assigned a label indicating whether it belonged to the tobacco plant or not. This labeled dataset is then utilized to train the SVM classifier, enabling it to distinguish between plant and non-plant descriptors. In order to estimate the total plant count, the tobacco plants are further grouped into clusters using the Euclidean distance. This clustering process grouped similar plant descriptors together, forming distinct clusters. The number of centroids, which represent the centers of these clusters, was then evaluated to determine the total plant count. The results obtained using the SVM-based approach are not satisfactory, as explained in further chapters of our study. Consequently, we explored an alternative approach by evaluating a Deep Learning (DL) semantic segmentation architecture.

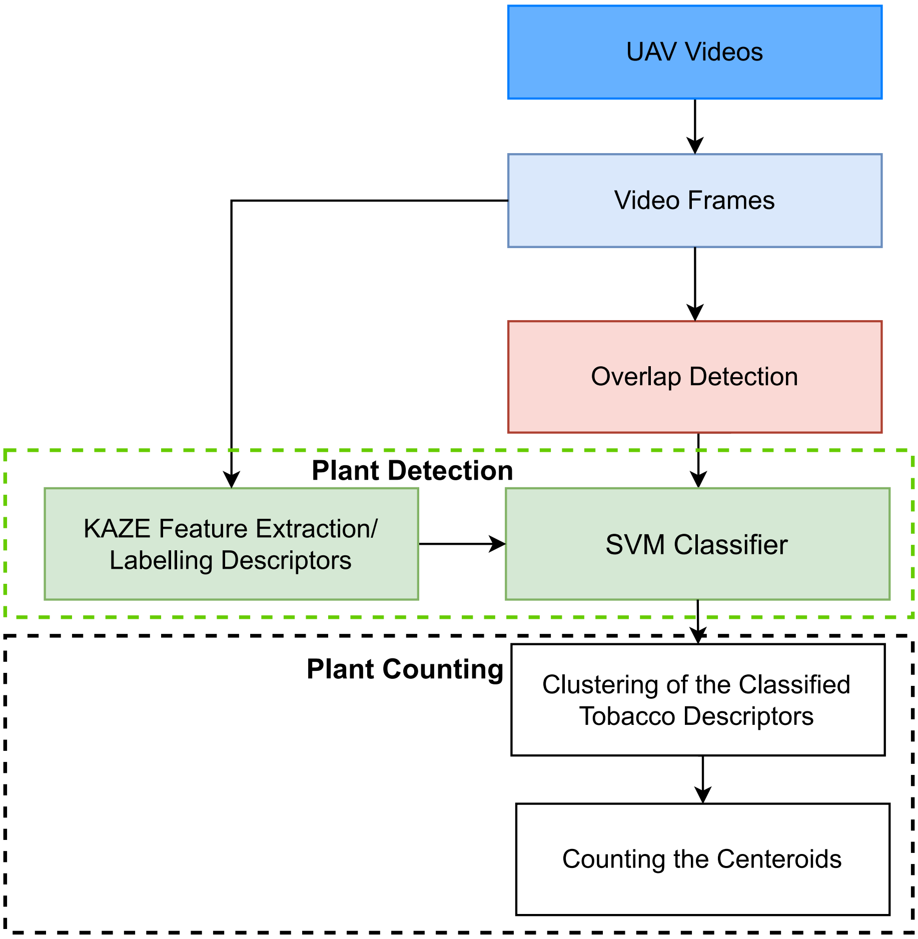


Figure 3.7: Proposed general framework using SVM

3.4 Semantic Segmentation as Detection Module

Segmentation is a crucial technique in various applications, including object detection and autonomous vehicles. Semantic segmentation is a technique that assigns labels to each pixel in an image, thereby segmenting the entire scene. In this research, the authors used LabelMe[26] to annotate 97 images for training, validation, and testing a deep learning architecture U-Net [27] for semantic segmentation.

3.4.1 U-Net

For image segmentation, a particular kind of network is employed: the UNet architecture. The U-Net model is a type of convolutional neural network that can extract features from low-resolution and small-sized images. It has a U-shaped structure and is made up of a bridge connecting a decoder network (expanding path) and an encoder network (contracting path).

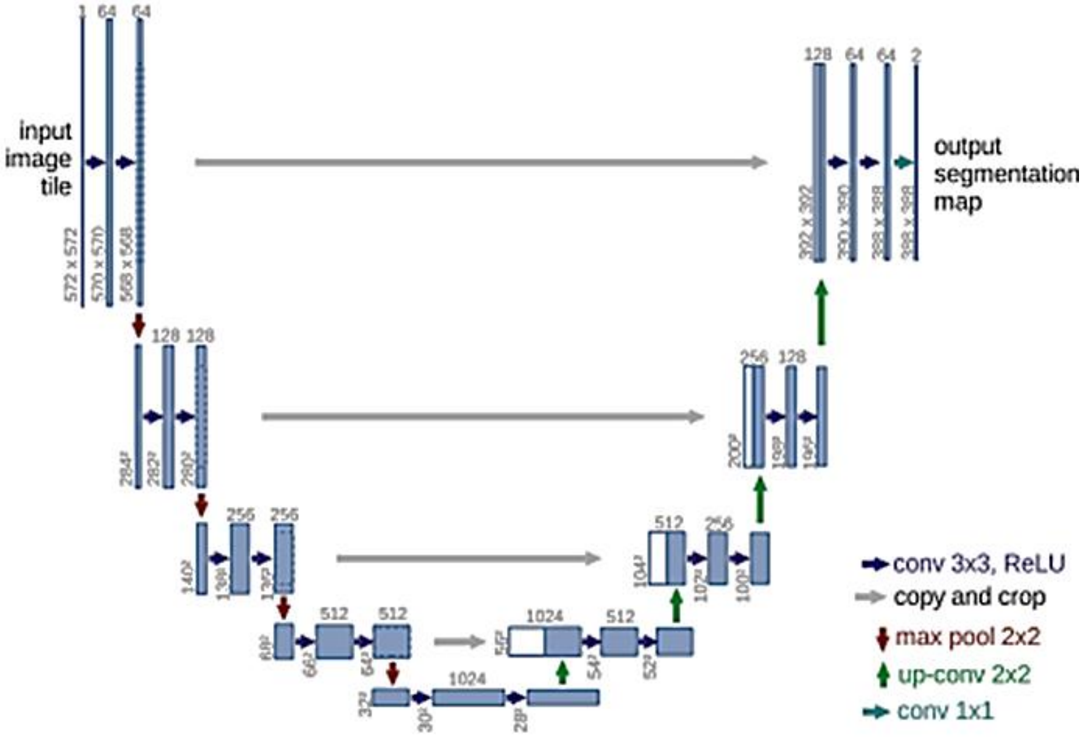


Figure 3.8: U-Net architecture[27]

The encoder network serves as a feature extractor, learning an abstract representation of the input image via a series of encoder blocks. Each encoder block typically consists of two consecutive operations: a convolutional layer followed by a non-linear activation function such as ReLU (Rectified Linear Unit) and a max pooling layer. Decoder consists of a series of decoder blocks. Each decoder block typically consists of an upsampling or transposed convolutional layer to increase the spatial dimensions while reducing the number of feature channels. This is followed by a concatenation operation that combines feature maps from the corresponding encoder block through skip connections as shown in Figure 3.8. It has been found to outperform other popular architectures like SegNet, PSPNet, and DeepLab v3+ in comparative studies[28]. U-Net has become popular in agriculture domain and has become popular in agriculture for plant and weed segmentation and classification[29].

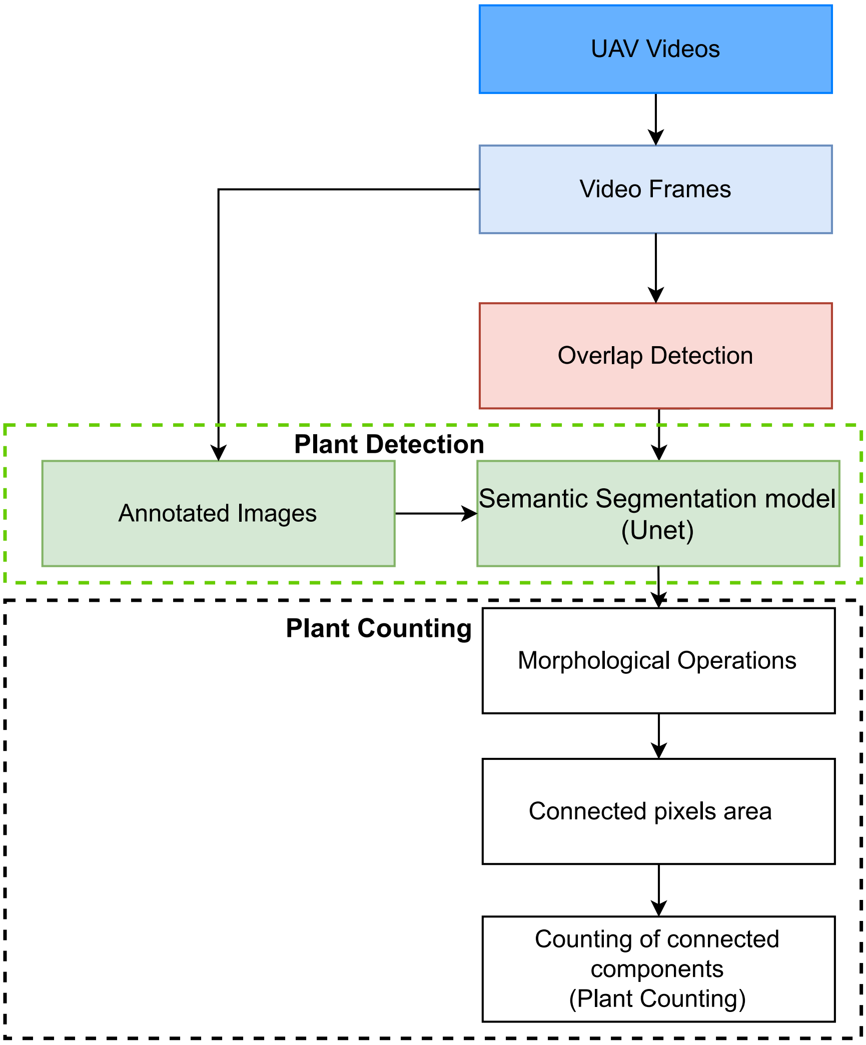


Figure 3.9: Proposed framework with semantic segmentation for plant detection

Encoder of the U-Net can be pretrained network, commonly used pretrained networks include VGG16, RESNET34, and Inceptionv3.

3.4.1.1 VGG16

VGG16 is a convolutional network [30] shown in Figure 3.9. It is trained on ImageNet dataset, it is the improved version of AlexNet. VGG16 consists of a total of 138 million parameters. A notable aspect of the architecture is that all convolutional kernels have a size of 3x3, while the max-pooling kernels have a size of 2x2 with a stride of two.

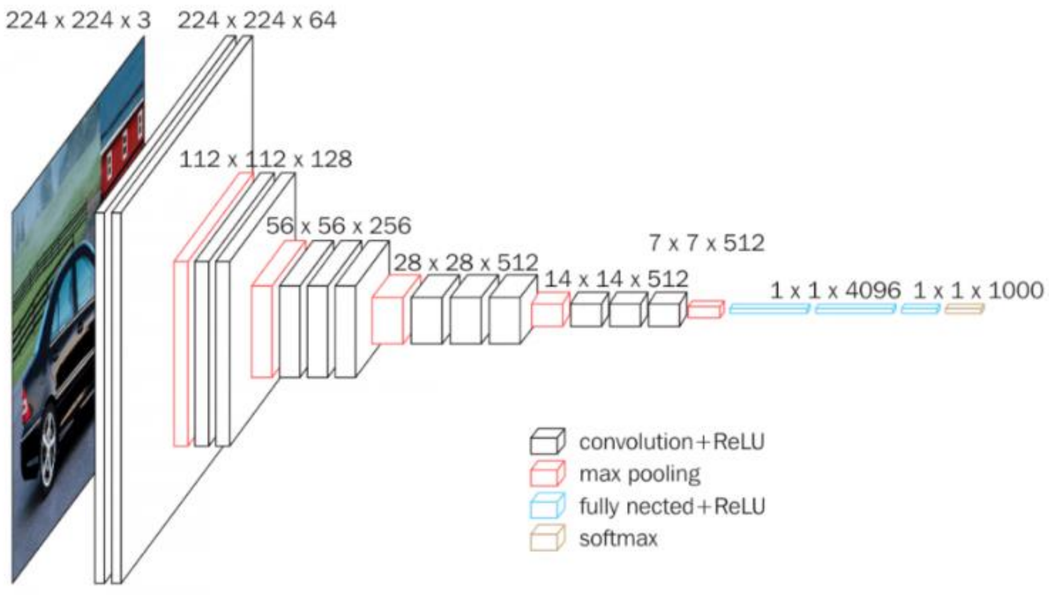


Figure 3.10: VGG16 Architecture[27]

3.4.1.2 ResNet34

Deep networks are not always the better they might result in vanishing gradient problem. Residual Networks known as ResNets[31] effectively solve this vanishing gradient problem. ResNets are composed of residual blocks, ResNet34 a 34-layer residual neural network is shown in Figure 3.11. ResNet presents a novel approach to address the vanishing gradient problem by introducing "skip connections" as an innovative solution. This technique involves stacking multiple identity mappings, which are convolutional layers that initially have no impact on the output, and then skipping these layers while reusing the activations from the preceding layer. By implementing skip connections, ResNet accelerates the initial training process by compressing the network into a reduced number of layers.

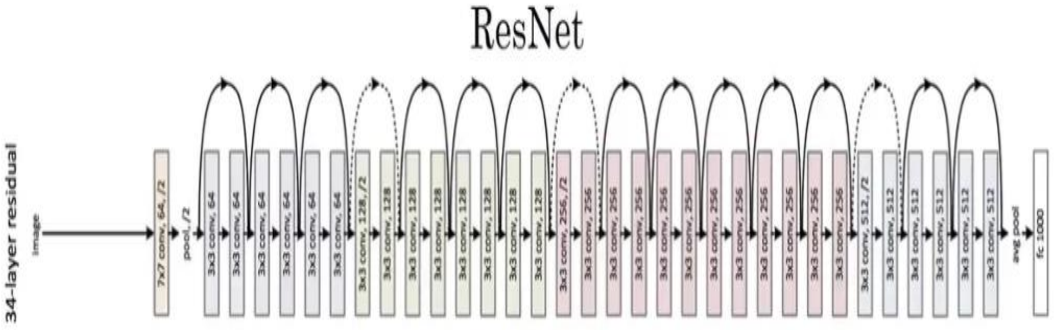


Figure 3.11: ResNet34 Architecture[31]

3.4.1.3 InceptionV3

When compared to VGGNet, Inception Networks[32] have demonstrated higher computational efficiency, both in terms of the network's parameter count and the economical cost incurred, including memory and other resource utilization. By incorporating factorizing convolutions and aggressive dimension reductions within a neural network, they have achieved comparatively lower computational costs while preserving high quality.

3.4.2 Proposed Framework with U-Net

We have evaluated U-Net as plant detection module in the proposed general framework as shown in Figure 3.9. The U-Net with VGG16 as its encoder was trained using Google Colab[33]. The segmentation results in binary images where vegetation pixels are assigned one, and non-vegetation pixels are assigned zero. Median blurring and morphological operations are used to remove noise, and counting is performed using connected pixel area to obtain the total tobacco plant count in the images. However, counting plants that appear on the boundary of the images leads to inaccurate total counts. Therefore, corner plants in one frame are counted and ignored in the next frame, resulting in better accuracy but missing some plants. To address this, the authors evaluated object detection performance for plant counting.

3.5 Object Detection as Detection Module

State-of-the-art object detection models have made significant advancements in the field, with one prominent example being YOLO (You Look Only Once) [8], [18]. In our study, we specifically focused on evaluating the performance of YOLOv7 [18] for the task of plant counting. YOLOv7 is a variant that builds upon the YOLO architecture, incorporating various improvements and optimizations. By utilizing YOLOv7 for plant counting, we aimed to assess its capabilities and effectiveness in accurately detecting and counting plants in an efficient manner. This involved analyzing its performance on a specific dataset and evaluating its accuracy, speed, and robustness in comparison to other object detection models.

The choice of YOLOv7 for this study was driven by its reputation as a reliable and high-performing object detection model. By evaluating its performance specifically in the context of plant counting, we sought to contribute to the understanding of its applicability in agricultural and plant-related applications.

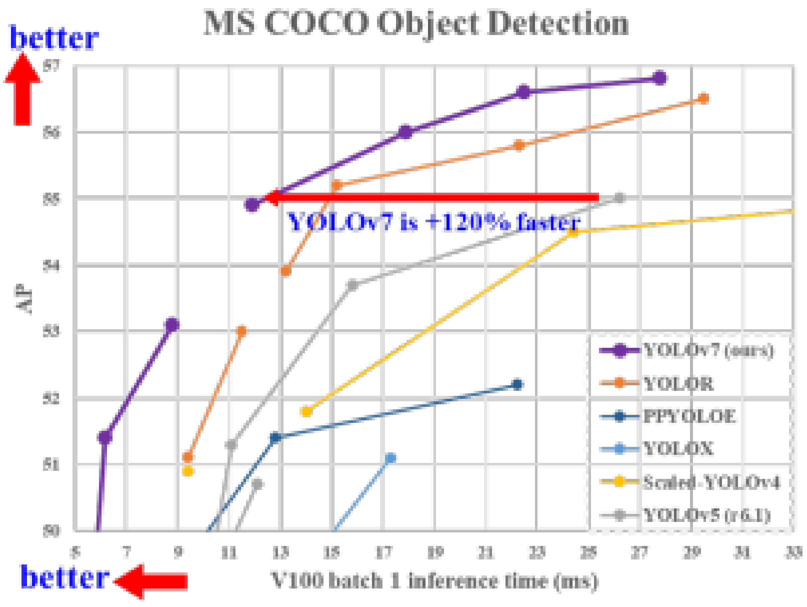


Figure 3.12: Comparison of other object detection models with YOLOv7[34]

3.5.1 YOLOv7

YOLOv7 is optimized with model reparameterization, compound model scaling and dynamic label assignment without increasing the inference cost. It outperformed transformer-based object detection models. Its comparison with other models is shown in Figure 3.12.

3.5.2 Proposed Framework with YOLOv7

Pre-trained basic YOLOv7 is used instead of its scaled version. The model is trained on 83 images and annotated using LabelImg [35]. After overlap detection, the frames are passed to train YOLO, and then number of bounding boxes are calculated to find the total count as shown in Figure 3.13.

The highest accuracy is achieved at a confidence threshold of 0.6. The problem of corner plants is much less than semantic segmentation as a bounding box is formed after correctly classifying something as tobacco rather than on connected pixels area. The above-mentioned techniques require very little time compared to state-of-the-art orthomosaic-based counting techniques. To make the system real-time, we evaluate an object detection model with tracking for tobacco plant counting.

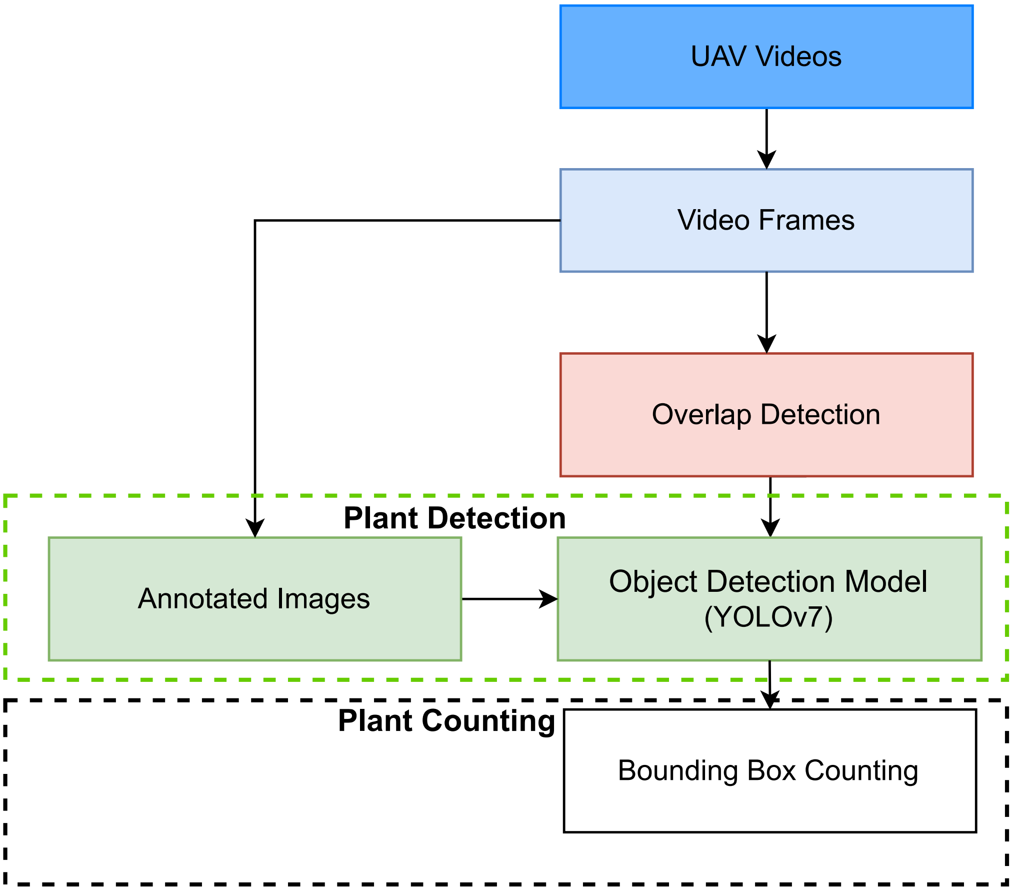


Figure 3.13: Proposed framework with object detection

3.6 Object Detection and Tracking

We have assessed the effectiveness of object detection combined with tracking for plant counting. This approach eliminates the need for overlap detection, as the plants are tracked, which prevents multiple counts of the same plant due to overlap. We conducted evaluations using recorded videos, but this technique can also be applied in real-time scenarios. For real-time plant counting, we integrated the YOLOV7 trained in the previous section with the SORT[19] algorithm. See Figure 3.14 for a graphical representation of this approach.

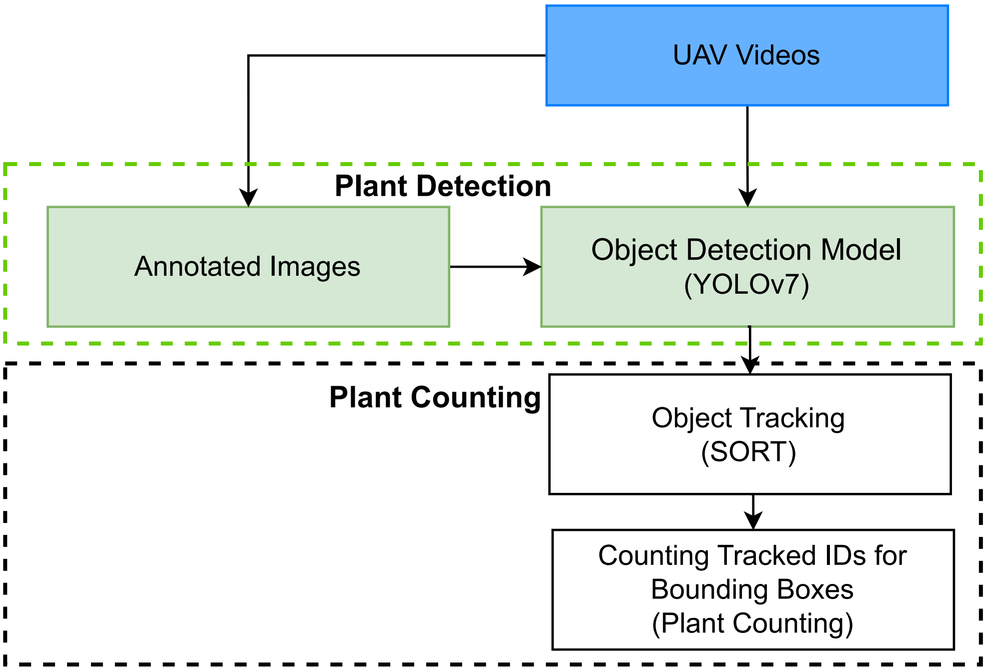


Figure 3.14: Proposed framework with object detection & tracking

Chapter 4: Results & Discussion

In order to assess the efficiency of the general plant counting development, it is crucial to critically evaluate each module. We conducted a comprehensive evaluation of every module of to gain valuable insights into the proposed methods.

4.1 Proposed Framework with Semantic Segmentation

In the evaluation of the proposed framework with semantic segmentation, we have evaluated plant detection module and the overall counting independently.

4.1.1 Evaluation Metrics for Semantic Segmentation

Evaluation of semantic segmentation model, involve several such as Intersection over Union (IOU), Precision, and Recall. These metrics play a crucial role in assessing the performance and accuracy of the model's predictions.

4.1.1.1 IOU

IOU is a widely utilized metric for semantic segmentation, ranging from 0 to 1. A value of 0 indicates the worst-case scenario, where there is almost no overlap between the predicted object and the ground truth. Conversely, a value of 1 represents a perfect prediction with 100% overlap between the ground truth and the prediction.

$$IOU = \frac{Area\ of\ Union}{Area\ of\ Overlap} \qquad (1)$$

4.1.1.2 Precision

Precision measures the proportion of true positives (TP) to the total number of positive predictions (TP + FP). A high precision value indicates accurate predictions with a low rate of false positives. It quantifies the model's ability to correctly classify positive instances.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

4.1.1.3 Recall

Recall, on the other hand, calculates the proportion of true positives (TP) to the total number of actual positive cases (TP + false negatives (FN)). A high recall value indicates that the model predicts most positive cases as positive. It assesses the model's ability to capture most positive instances.

TP: the number of correctly classified pixels belonging to the target class.

FP: the number of pixels incorrectly classified as the target class when they are not

FN: represents mistakenly classified pixels of the target class as other classes or background.

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

We conducted an evaluation of the U-Net architecture using **Eq. 1**, **Eq. 2**, and **Eq. 3**. This evaluation encompassed two approaches: transfer learning and fine-tuning.

4.1.2 Transfer Learning VS Fine Tuning

Transfer learning involved initializing the network with pre-trained weights and freezing all layers except the fully connected layers. On the other hand, fine-tuning included retraining all layers of the network. We have evaluated transfer learning and fine tuning for different backbones as shown in Figure 4.1.

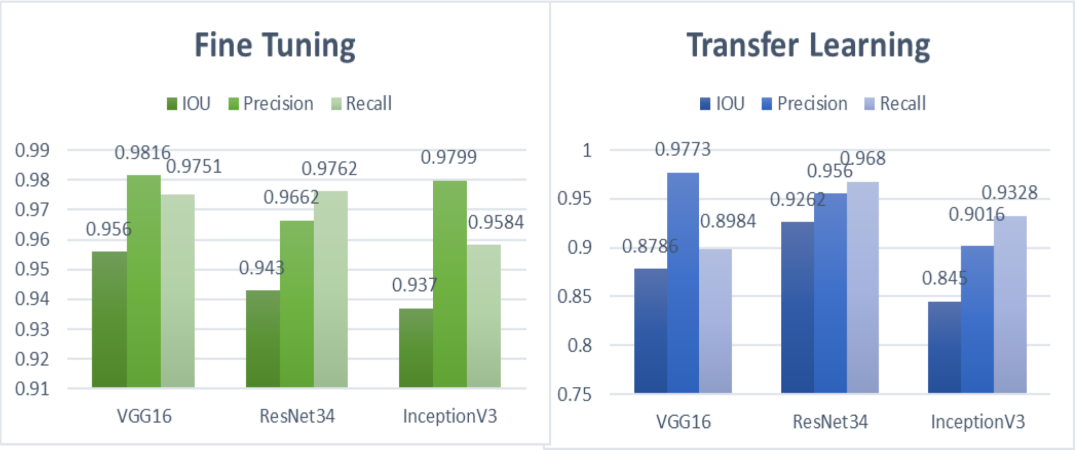
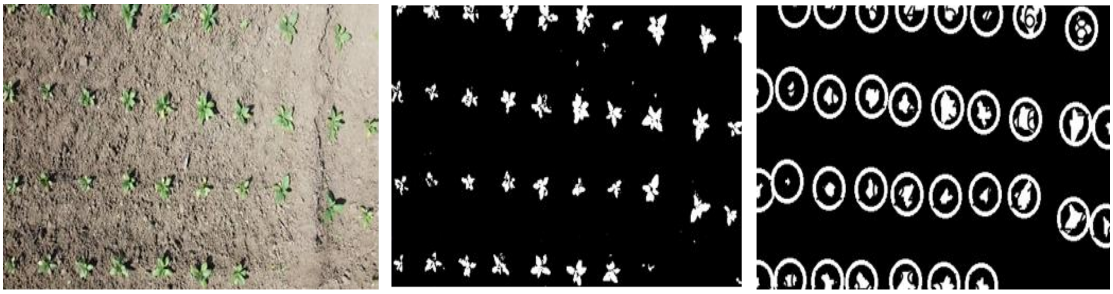


Figure 4.1: Comparison of fine-tuning and transfer learning with different backbone models for the U-Net architecture

The results clearly indicate that fine-tuning produces better outcomes for RESNET34, VGG16, and Inception V3 in our specific problem. Fine-tuning, which involves further training of pre-trained models, leads to improved accuracy, particularly when dealing with smaller datasets. This advantage makes the model more applicable to other plant species even with limited labeled data.

Among the different configurations, U-Net with a fine-tuned VGG16 architecture as the backbone model achieves the highest IOU score of 0.9556. This score signifies the model's strong performance in accurately predicting the overlap between the predicted and ground truth segmentation masks. Moreover, this configuration also demonstrates higher precision, indicating the model's capability to minimize false positive predictions. Input image, U-NET prediction, and counting after morphological operations are shown for the initial/first frame with no overlap masked region in Figure 4.2a, Figure 4.2b, and Figure 4.2c respectively.



(a) Input Frame (b)Prediction (c) Counting results

Figure 4.2: Visualization of results obtained using U-Net for detection

4.1.3 Evaluation Metrics for Plant Counting

Proposed General framework is evaluated using the following metrics:

4.1.3.1 Precision

.A precision score of 1 signifies a high level of accuracy and reliability, indicating that all identified plants can be assumed to be real.

$$Precision = \frac{No.\ of\ correctly\ Identified\ Plants}{No. of\ correctly\ identified\ plants\ +\ No. of\ incorrectly\ identified\ plants} \quad (4)$$

4.1.3.2 Recall

A recall score of 1 indicates high sensitivity, capable of detecting all plants present in the field.

$$Recall = \frac{No.\ of\ Correctly\ Identified\ Plants}{(No.\ of\ correctly\ identified\ plants\ +\ Number\ of\ missed\ plants)} \quad (5)$$

4.1.3.3 F1 Score

An F1 score of 1 indicates accurate identification and counting of all plants (high recall) while minimizing false positives (high precision). This represents an ideal scenario without any overlooked plants or misidentifications during the counting procedure.

$$F1\ Score = \frac{2 \times Precision\ \times Recall}{Precision + Recall} \qquad (6)$$

Table 4.1: Crop Emergence Estimation using Semantic Segmentation

Name	No. of images	True Count	Estimated	Precision	Recall	F1 Score
DATASET-1	28	472	428	1	0.906	0.951
DATASET-2	29	478	426	1	0.89	0.943
DATASET-3	32	522	510	1	0.977	0.988
DATASET-4	40	536	464	1	0.86	0.930
DATASET-5	61	540	463	1	0.857	0.923
Average				1	0.898	0.947

Table 4.1 displays the estimated number of plants based on semantic segmentation. Five different video clips of data from three tobacco fields are used to evaluate the model. Dataset 1, Dataset 2, Dataset 3, Dataset 4, and Dataset 5 are the different video clips. The estimated count is the outcome of the framework, and the ground truth is a manual count performed in a video. The precision, recall, and F1 score are calculated using Eq. 4, Eq. 5, and Eq. 6 respectively.

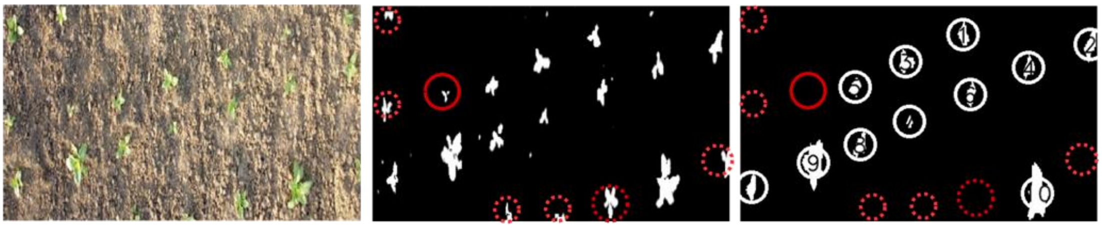
The proposed framework achieved impressive results with an average F1 score of 0.947, a precision score of 1, and a recall score of 0.8992. These metrics indicate the framework's ability to accurately segment and count objects.

One notable observation is that a higher precision in the semantic model leads to a higher precision in the counting process, effectively minimizing false positive detections. Additionally, a higher Intersection over Union (IOU) of the U-Net model contributes to accurate segmentation masks, reducing the chances of both false positive and false negative plant detections. Consequently, the overall plant counting becomes more precise.

4.1.4 Challenges and Limitations of the Proposed Framework with Semantic Segmentation

Figure 4.3a represents the input image with the manually cropped overlap region, while Figure 4.3b shows the predicted image after the same region has been cropped. Additionally, Figure 4.3c displays the result of the proposed framework with semantic segmentation, which includes cropping the identified overlap region, applying morphological operations, and performing counting based on connected pixel area.

The data presented in Table 4.1 reveals a consistent pattern of underestimation in the estimated counts compared to the ground truth counts. This underestimation can be attributed to two primary factors depicted in Figure 4.3. Firstly, there are plants that are not included in the count due to their removal from the predicted images through the application of morphological operations. These operations may inadvertently eliminate certain plants, resulting in their exclusion from the final count. Secondly, we have excluded boundary plants in one of the two consecutive frames are excluded to avoid double counting plants that partially reside at the boundaries of both frames. Due to this some of the plants are not counted at all resulting in underestimation.



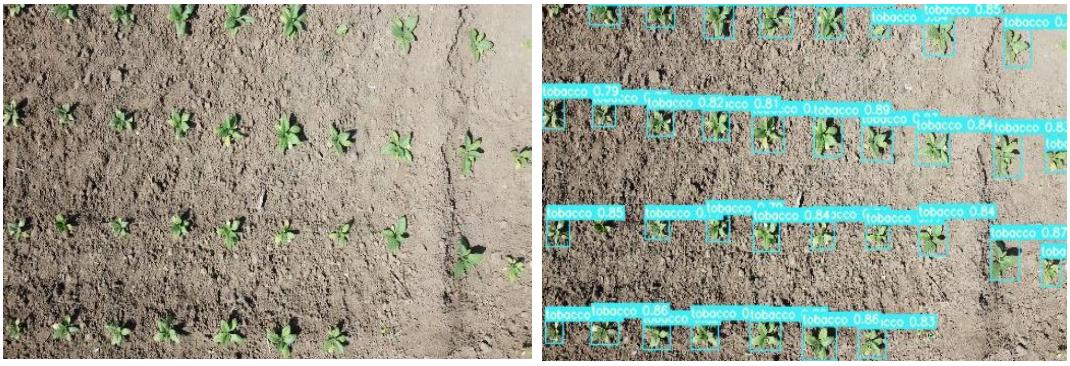
(a) Input (b)Plant detection (c)Plant counting

Figure 4.3: Qualitative analysis of proposed framework with U-Net

4.2 Proposed Framework with Object Detection

In this study, the model we used for plant counting was trained on a carefully curated dataset consisting of 83 images. These images were selected to represent a diverse range of plant types, growth stages, and environmental conditions commonly encountered in the target applications.

To ensure the reliability and generalizability of our model, we performed rigorous validation on a separate set of 20 images that were not included in the training dataset. This validation set allowed us to assess the model's performance on unseen data and determine its ability to generalize to new instances. The evaluation of the model on the validation set yielded exceptional results, with a mean average precision (mAP) of 0.988. This high mAP score is indicative of the model's remarkable accuracy and performance in accurately detecting plants. The tobacco detection using this trained object detection model (YOLOv7) is demonstrated in Figure 4.4. The counting of tobacco plants is accomplished by determining the number of bounding boxes generated after the detection process.



(a) Input Frame (b)Prediction

Figure 4.4: Visualization of results obtained using YOLOv7 for detection

Table 4.*2*: Crop Emergence Estimation using Object Detection

Name	No. of images	True Count	Estimated	Precision	Recall	F1 Score
DATASET-1	28	472	460	1	0.97	0.987
DATASET-2	29	478	500	0.956	1	0.977
DATASET-3	32	522	535	0.975	1	0.988
DATASET-4	40	536	498	1	0.929	0.963
DATASET-5	61	540	459	1	0.85	0.918
Average				0.986	0.9498	0.966

Table 4.2 presents the evaluation results of the proposed algorithm. The evaluation metrics include ground truth, accuracy, precision, and recall, which were calculated as described in the previous section. The algorithm achieved an average F1 score of 0.9667, a precision score of 0.9852, and a recall score of 0.9484. These results

demonstrate that the proposed algorithm is a reliable solution for plant detection, exhibiting good sensitivity.

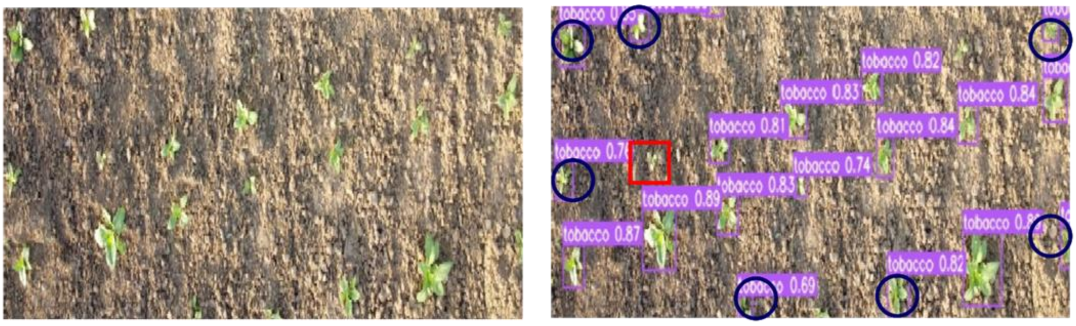
The achieved accuracy is higher than that of semantic segmentation alone, primarily due to the superior precision of the model employed in the proposed algorithm. The higher precision ensures the detection of all plants in the image. However, there is still some variance observed in the estimated counts compared to the ground truth counts.

4.2.1 Challenges and Limitations of the Proposed Framework with Object Detection

This variance in the observed estimated count can be attributed to two factors. Firstly, when plants are located at the corners of consecutive frames, they may be counted twice, leading to a higher estimated count than the ground truth. This occurs because partially present plants at the corners can appear as separate entities in both frames, resulting in their double counting.

Secondly, datasets that contain smaller plants tend to have a lower estimated count compared to the ground truth. This is due to the challenge of accurately identifying very small tobacco plants, resulting in their exclusion from the count.

To provide visual evidence of these factors, Figure 8 showcases the same input image as in Figure 4.2a. In Figure 4.5, the solid blue circles represent the partially present corner plants, highlighting the possibility of double counting. The figure also emphasizes the presence of smaller plants that are not classified as tobacco, contributing to the variance between the estimated count and the ground truth.

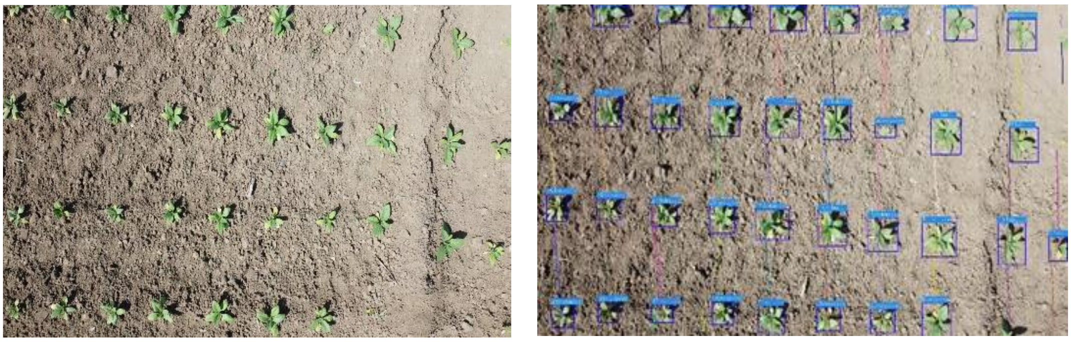


(a) Input Frame (b) Plant detection & counting

Figure 4.5: Qualitative analysis of proposed framework with YOLOv7

4.3 Proposed Framework with Object Detection and Tracking

The process of tobacco plant estimation utilizing object detection and tracking is illustrated in Figure 4.6, providing a visual representation of the algorithm's workflow. The detected tobacco plants are enclosed within bounding boxes, which are annotated with unique IDs for identification and tracking purposes. The estimation of the plant count is derived from the total number of generated IDs, allowing for an accurate quantification of the tobacco plants in the given video clips. To assess the performance and effectiveness of this technique, comprehensive evaluation results are presented in Table 4.3. This table showcases the quantitative metrics obtained from the evaluation, offering insights into the algorithm's performance and its ability to accurately estimate the plant count in the video clips.



(a) Input frame

(b) Detection & tracking

Figure 4.6: Visualization of results obtained using YOLOv7 & SORT

The proposed methodology demonstrates great promise for real-time estimation, showcasing a recall score of 1.

Table 4.3: Crop Emergence Estimation using Real-Time Applicable Framework

Name	True Count	Estimated	Precision	Recall	F1 score
DATASET-1	472	496	0.951	1	0.974
DATASET-2	478	518	0.922	1	0.959
DATASET-3	522	566	0.922	1	0.959
DATASET-4	536	560	0.957	1	0.978
DATASET-5	540	578	0.931	1	0.966
Average			0.9336	1	0.967

The high recall score indicates excellent sensitivity in plant detection, as it successfully captures all the plants present in the scene. Furthermore, the methodology achieves a good enough precision, suggesting a high level of accuracy in the counting process. The combination of high recall and satisfactory precision makes this approach a promising solution for plant counting tasks. By effectively detecting and accurately counting the plants, it addresses the key objectives of the estimation process. The methodology's ability to achieve a recall score of 1 implies that it leaves no room for missing any plants, ensuring comprehensive coverage.

With its real-time capabilities and a balance between sensitivity and precision, this methodology offers significant potential in various applications where accurate and timely plant estimation is crucial. Researchers and practitioners can leverage this promising approach to enhance plant monitoring, agricultural management, and related fields where counting and tracking plants are essential for decision-making and analysis.

4.3.1 Challenges and Limitations of the Proposed Framework with Object Detection & Tracking

An observation from the evaluation is that the estimated count is higher than the true value. This discrepancy can be attributed to the switching IDs of partial tobacco plants at the side rows of the video clips. When occlusion occurs, and a partially visible plant reappears in the frame, it is detected as a new entry with a different ID. As a result, the count is inflated, leading to an overestimation of the tobacco plant population.

This issue highlights a challenge in accurately tracking and counting plants when dealing with occlusion and the reappearance of the same plant. The switching IDs phenomenon can introduce inaccuracies in the estimation process, especially in scenarios where partial plants are present at the edges of the video clips. Addressing this challenge requires further refinement of the object tracking algorithm to handle occlusion and maintain consistent IDs for the same plant throughout the video sequence. By mitigating the switching IDs problem, more accurate and reliable plant counts can be obtained using the object detection and tracking approach.

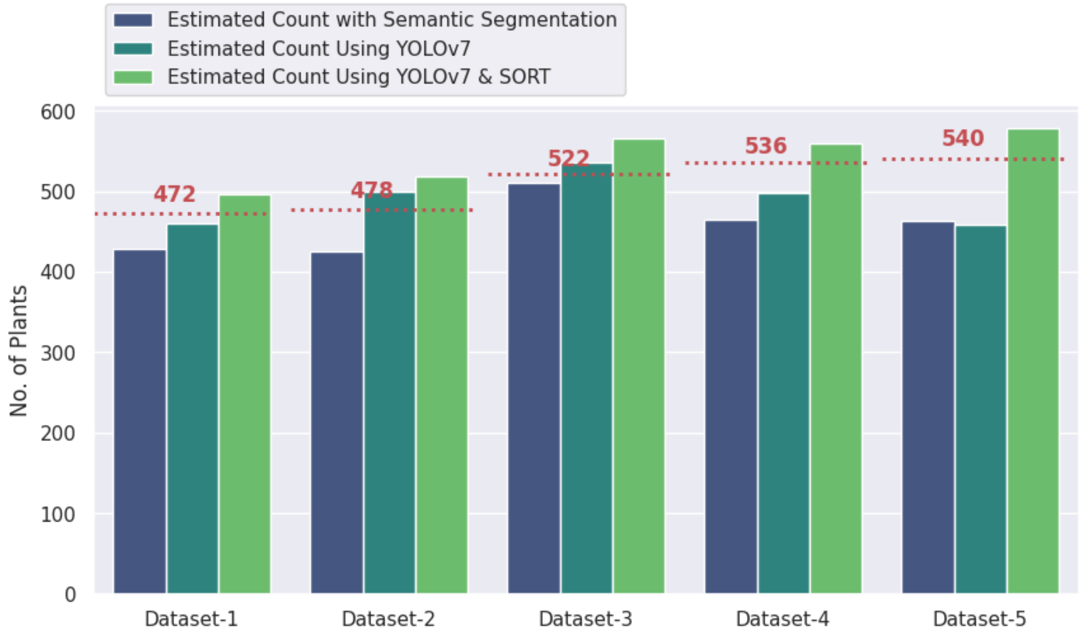


Figure 4.7: Comparison of plant counting techniques

Crop emergence estimation with all the techniques is observed in Figure 4.7. which presents a detailed comparison of the proposed techniques, showcasing notable patterns in the estimated plant counts. The findings demonstrate distinct characteristics among the different approaches.

The semantic segmentation technique consistently yields lower plant counts, suggesting a tendency towards underestimation. On the other hand, the real-time plant counting framework consistently produces higher counts, indicating a tendency towards overestimation. This consistent pattern observed in both techniques suggests that their performance is independent of the nature of the dataset.

In contrast, the object detection-based technique exhibits variance in the plant count. This variance indicates a dependency on the specific characteristics of the dataset, such as plant density, size, and arrangement. The results highlight the sensitivity of the object detection approach to the dataset's unique features, leading to fluctuations in the estimated plant counts.

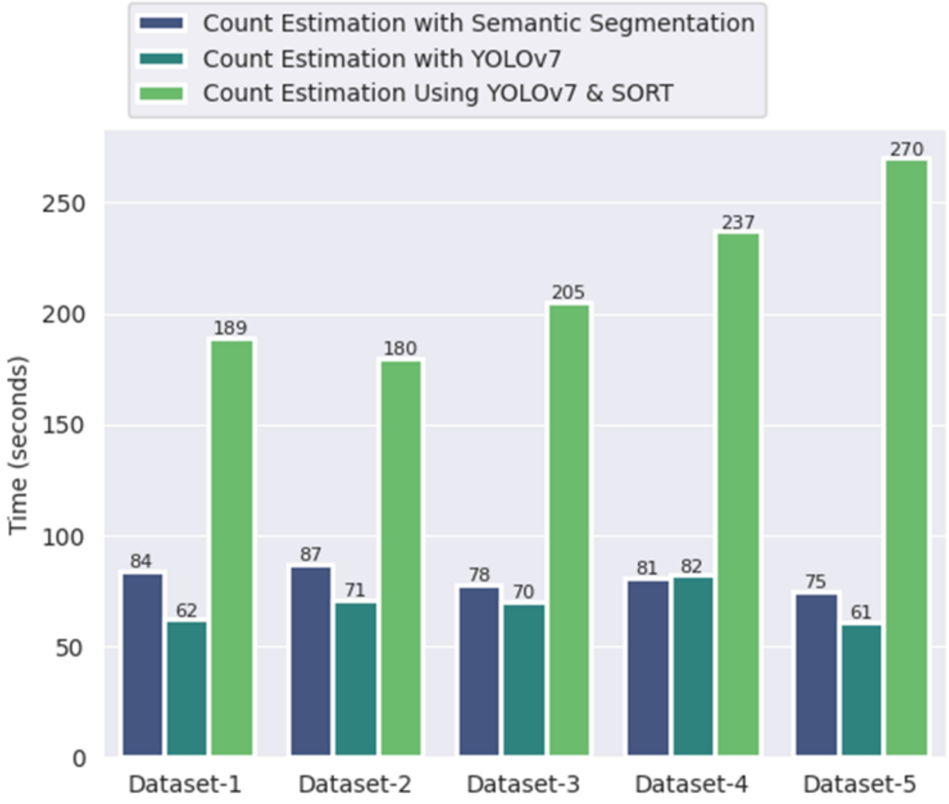


Figure 4.8: Speed Analysis

Figure 4.8 illustrates the speed analysis of the proposed techniques. The real-time plant counting framework, evaluated on video data, demonstrates relatively higher average processing time compared to the other methods. However, the results highlight the efficiency of the proposed pipelines, as depicted in Figure 1, which require significantly less time and computational power compared to state-of-the-art techniques that involve orthomosaic formation.

It is worth noting that all the computations for this research were performed using Google Colab. This choice of computing platform contributes to the efficient processing of the proposed techniques while maintaining computational feasibility.

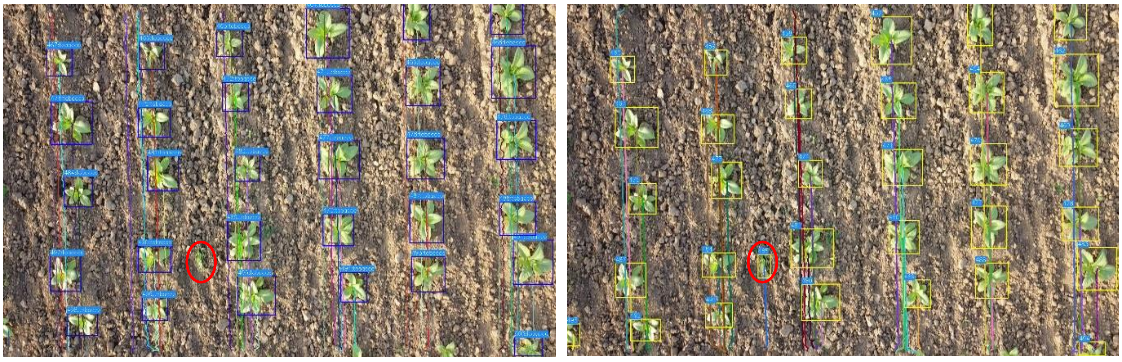
The findings presented in Figure 10 provide valuable insights into the trade-off between speed and accuracy in the various plant counting approaches. Researchers and practitioners can utilize this information to select the most suitable technique based on their specific requirements, considering the desired level of real-time capability and computational resources available.

4.3.2 Performance Comparison of YOLOv7 and YOLOv8 for Plant Counting in the Proposed Framework

We have also evaluated the recently updated version of YOLO, YOLOv8 for our proposed framework that incorporates object detection and tracking. The detection module using YOLOv7 was replaced with YOLOv8.

For better evaluation we have trained and validated YOLOv8 with the same images as used for training and validating YOLOv7 that results in 0.99 MAP.

Figure 4.9 shows tobacco plant detection with both YOLOv7 and YOLOv8 on the same video clip. Solid red circle shows weed. It can be clearly observed that YOLOv8 classified weed as tobacco plant. Whereas YOLOv7 did not detected weed as tobacco plant.

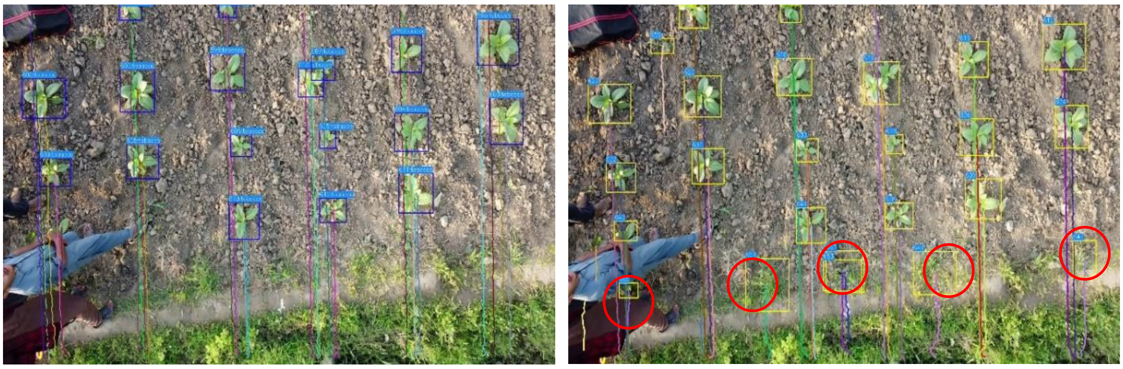


(a) UsingYOLOv7 & SORT (b) YOLO8 & SORT

Figure 4.9: Comparison of plant detection with YOLOv7 vs YOLOv8

Figure 4.9 illustrates the occurrence of misclassifications where weeds are incorrectly identified as tobacco plants within our proposed framework. The solid red circles highlight all the instances of misclassification. It is evident that the performance of the YOLOv8 model is not up to the mark for our proposed general framework.

To address this issue, one potential solution is to enhance the training of the model by incorporating three distinct classes: weeds, tobacco plants, and ground. By including a specific class for weeds, the model can learn to differentiate more effectively between the different types of vegetation present in the field. This approach has the potential to improve the accuracy of the plant classification process within our framework. Misclassification of weed as tobacco plant is represented in Figure 4.10 in which solid red circle highlights all the misclassified instances. Hence, for our proposed framework YOLOv8 shows poor results. This might be improved by training the model with three classes; weed, tobacco, and ground.

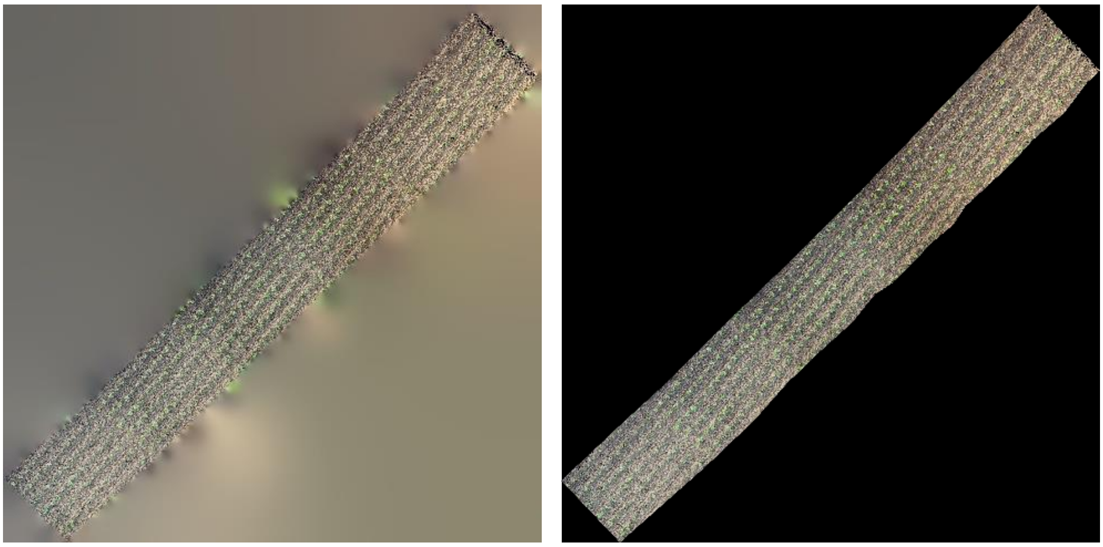


(a) Using YOLOv7 & SORT (b) Using YOLOv8 &SORT

Figure 4.10: Comparative analysis of plant detection performance between YOLOv7 & YOLOv8 models

4.4 Orthomosaic based Plant Counting

We also tested orthomosaic-based counting to highlight the time-efficiency of the proposed plant counting framework, which consists of orthomosaic formation followed by patchifying (extraction of small images) this very high-resolution image to smaller images, plant detection on patches using previously discussed in Section 4.2 trained YOLOv7, and reconstruction of large image from these plant-detected patches.

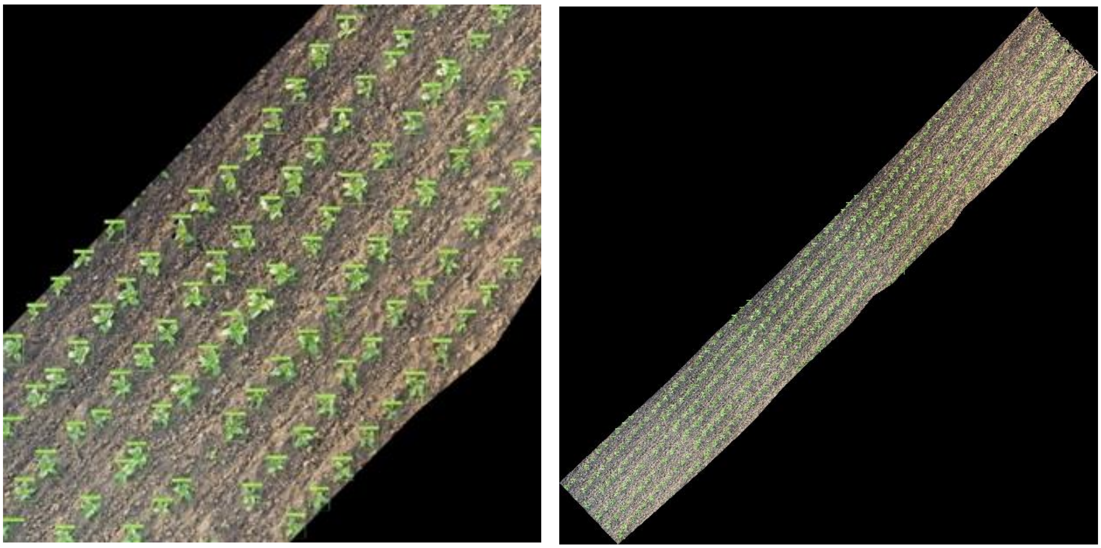


(a) Orthomosaic of Dataset-1 (b) Blurred Edges Removed

Figure 4.11: Orthomosaic Formation

Figure 4.11 depicts the orthomosaic of Dataset-1. Agisoft Metashape[36] is used to create the orthomosaic. Orthomosaic is preprocessed to remove blurred edges as shown in Figure 4.11b and resized for patch extraction with the least amount of overlap between patches. The reconstruction of a large image from detected patches is used to identify plants that are counted twice because they are present in more than one patch.

During our evaluation, we have observed a limitation in the orthomosaic formation process, which is the possibility of missing frames. In some instances, certain frames not included during the construction of the orthomosaic in the alignment o frames. As a result, these excluded frames can lead to an underestimation in the manual plant count within the orthomosaic. It is important to address this limitation because the absence of these frames can result in an incomplete representation of the plant population within the area of interest. By missing frames, there is a risk of not accounting for all the plants present, leading to a lower count than the actual number of plants.



(a) Object detection on a patch (b) Recontruction after detection

Figure 4.12: Orthomosaic based Plant Counting

Figure 4.12 shows the plant detected patched of resolution 1920×1920 and a reconstructed large image to detect the double detection of a single plant present in more than one patch. Figure 4.13 depicts that manually counting plants in an orthomosaic results in a count of 457 which is lower than the observed ground truth count of 472 from a video clip. This underestimation suggests that one or more frames were missed during the time-consuming process of orthomosaic formation, and it may also be attributed to blurred corners that hindered accurate plant identification. In contrast, the proposed object detection-based technique is also evaluated on patches extracted from the orthomosaic. The evaluation revealed an underestimation in plant count, with a total of 421 plants detected. This count is lower compared to the count achieved by the proposed framework.

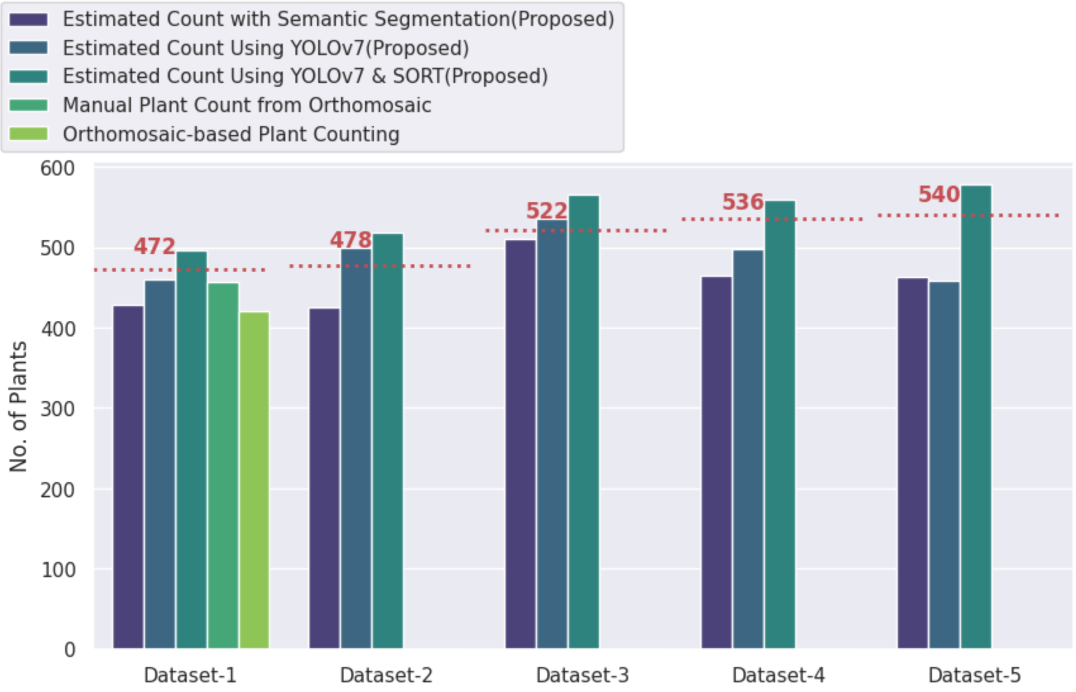


Figure 4.13: Comparison of proposed techniques with orthomosaic based plant counting

To highlight the time efficiency of the proposed framework we have done time comparison. The orthomosaic formation process for DATASET-1, as depicted in Figure 4.10a, required a significant amount of time, taking approximately 6639 seconds (equivalent to 1 hour and 50 minutes). In contrast, the proposed overlap detection algorithm, utilizing the same dataset and hardware specifications, achieved the desired results in a significantly shorter duration, completing the task in just 184 seconds. Both process are done on the same system with hardware specifications of 7.92 GB RAM, CPU Intel(R) Core(TM) i5-7200U, CPU 2.50GHz and AMD Radeon(TM) R5 M430 (Hainan) GPU.

Chapter 5: Conclusion

We conducted a comprehensive comparison of techniques for early stand count estimation of tobacco plants using UAV data and DL models. The primary objective was to find a more efficient and accurate approach for tobacco plant counting, considering both time efficiency and computational requirements.

The proposed generalized framework with different detection modules showcased promising results, demonstrating their potential for estimating tobacco plant counts in a shorter amount of time and with reduced computational power. One notable contribution of this study was the introduction of a novel approach for overlap detection based on visual features. This approach proved to be effective in achieving comparable counting accuracies to the computationally expensive orthomosaic-based methods commonly used in the field.

By combining overlap detection with semantic detection, the researchers observed variances in the results, with an average F1 score of 0.947. It was observed misclassified pixels could introduce noise to the binary images, leading to slight discrepancies in the results. To overcome this, the researchers further explored the combination of overlap detection with object detection, resulting in an improved average F1 score of 0.9667, surpassing the performance of the segmentation-based approach.

To evaluate the proposed techniques in real-time scenarios, a dedicated system for plant counting through object detection with tracking was assessed using recorded data. The system showed an average F1 score of 0.9672, exhibiting minimal variance.

While the techniques showed overall effectiveness, certain challenges and areas for improvement were identified. For instance, during overlap detection, plants located at the boundaries of the images were partially present in both frames, leading to potential recounting issues. To mitigate this, we opted to ignore boundary objects in subsequent frames for the semantic segmentation-based approach. But this leads missing plants in counting and results in underestimated count.

Moreover, the object detection approach exhibited higher sensitivity, resulting in the inclusion of some larger weeds being counted as plants and recounting some corner plants present in consecutive frames. We aim to address these by training the detection model on weed class data as well and feature matching of every detected tobacco at the end of the proposed pipelined framework will solve double counting of corner plants.

Additionally, the detection and tracking approach encountered id switching for partial plants located at the corners of the video. This issue will be resolved by ensuring the drone's position stability during data collection, preventing horizontal shifts of plants at the vertical corners.

However, it should be acknowledged that the collected aerial data used in the study had minimal weed infestation. So, the algorithm's performance might differ when dealing with higher levels of weed infestation. To address this limitation, future work will involve integrating the proposed algorithm with a weed classification technique previously proposed in [29]. Moreover all the datasets have one directional motion of the UAV.

We will evaluate the proposed methods on other plant species in the future, expanding our applicability beyond tobacco. This will enable a broader understanding of the techniques' effectiveness and adaptability across various crops.

The findings provide a foundation for enhancing precision agriculture practices and hold potential for broader applications in crop management and monitoring.

References

[1] D. Rahmawati, R. Alfita, M. Ulum, and D. Murdianto, “Tobacco Farming Mapping To Determine The Number Of Plants Using Contour Detection Method,” in *E3S Web of Conferences*, EDP Sciences, Dec. 2021. doi: 10.1051/e3sconf/202132804007.

[2] Z. Fan, J. Lu, M. Gong, H. Xie, and E. D. Goodman, “Automatic Tobacco Plant Detection in UAV Images via Deep Neural Networks,” *IEEE J Sel Top Appl Earth Obs Remote Sens*, vol. 11, no. 3, pp. 876–887, Mar. 2018, doi: 10.1109/JSTARS.2018.2793849.

[3] Y. Wang, Z. Zhou, D. Huang, T. Zhang, and W. Zhang, “Identifying and Counting Tobacco Plants in Fragmented Terrains Based on Unmanned Aerial Vehicle Images in Beipanjiang, China,” *Sustainability (Switzerland)*, vol. 14, no. 13, Jul. 2022, doi: 10.3390/su14138151.

[4] T. Liu, W. Wu, W. Chen, C. Sun, X. Zhu, and W. Guo, “Automated image-processing for counting seedlings in a wheat field,” *Precis Agric*, vol. 17, no. 4, pp. 392–406, Aug. 2016, doi: 10.1007/s11119-015-9425-6.

[5] P. Randelović *et al.*, “Prediction of soybean plant density using a machine learning model and vegetation indices extracted from RGB images taken with a UAV,” *Agronomy*, vol. 10, no. 8, Aug. 2020, doi: 10.3390/agronomy10081108.

[6] S. Oh *et al.*, “Plant counting of cotton from UAS imagery using deep learning-based object detection framework,” *Remote Sens (Basel)*, vol. 12, no. 18, Sep. 2020, doi: 10.3390/RS12182981.

[7] Q. Liu, Y. Zhang, and G. Yang, “Small unopened cotton boll counting by detection with MRF-YOLO in the wild,” *Comput Electron Agric*, vol. 204, Jan. 2023, doi: 10.1016/j.compag.2022.107576.

[8] S. Khaki, N. Safaei, H. Pham, and L. Wang, “WheatNet: A lightweight convolutional neural network for high-throughput image-based wheat head detection and counting,” *Neurocomputing*, vol. 489, pp. 78–89, Jun. 2022, doi: 10.1016/j.neucom.2022.03.017.

[9] A. Feng, J. Zhou, E. Vories, and K. A. Sudduth, “Evaluation of cotton emergence using UAV-based imagery and deep learning,” *Comput Electron Agric*, vol. 177, no. August, p. 105711, 2020, doi: 10.1016/j.compag.2020.105711.

[10] Y. Pang *et al.*, “Improved crop row detection with deep neural network for early-season maize stand count in UAV imagery,” *Comput Electron Agric*, vol. 178, no. August, p. 105766, 2020, doi: 10.1016/j.compag.2020.105766.

[11] C. Miao, A. Pages, Z. Xu, E. Rodene, J. Yang, and J. C. Schnable, “Semantic Segmentation of Sorghum Using Hyperspectral Data Identifies Genetic Associations,” *Plant Phenomics*, vol. 2020, pp. 1–11, 2020, doi: 10.34133/2020/4216373.

[12] J. Valente, B. Sari, L. Kooistra, H. Kramer, and S. Mücher, “Automated crop plant counting from very high-resolution aerial imagery,” *Precis Agric*, vol. 21, no. 6, pp. 1366–1384, 2020, doi: 10.1007/s11119-020-09725-3.

[13] L. P. Osco *et al.*, “A CNN approach to simultaneously count plants and detect plantation-rows from UAV imagery,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 174, pp. 1–17, Apr. 2021, doi: 10.1016/j.isprsjprs.2021.01.024.

[14] X. Bai *et al.*, “Rice Plant Counting, Locating, and Sizing Method Based on High-Throughput UAV RGB Images,” *Plant Phenomics*, vol. 5, Jan. 2023, doi: 10.34133/plantphenomics.0020.

[15] C. Nee, L. S. Conway, J. Zhou, N. R. Kitchen, and K. A. Sudduth, “Early corn stand count of different cropping systems using UAV-imagery and deep learning,” *Comput Electron Agric*, vol. 186, no. May, p. 106214, 2021, doi: 10.1016/j.compag.2021.106214.

[16] B. T. Kitano, C. C. T. Mendes, A. R. Geus, H. C. Oliveira, and J. R. Souza, “Corn Plant Counting Using Deep Learning and UAV Images,” *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019, doi: 10.1109/lgrs.2019.2930549.

[17] H. Yang *et al.*, “Multi-object tracking using Deep SORT and modified CenterNet in cotton seedling counting,” *Comput Electron Agric*, vol. 202, Nov. 2022, doi: 10.1016/j.compag.2022.107339.

[18] N. Wojke, A. Bewley, and D. Paulus, "SIMPLE ONLINE AND REALTIME TRACKING WITH A DEEP ASSOCIATION METRIC."

[19] A. I. B. Parico and T. Ahamed, "Real time pear fruit detection and counting using yolov4 models and deep sort," *Sensors*, vol. 21, no. 14, Jul. 2021, doi: 10.3390/s21144803.

[20] C. Tan, C. Li, D. He, and H. Song, "Towards real-time tracking and counting of seedlings with a one-stage detector and optical flow," *Comput Electron Agric*, vol. 193, Feb. 2022, doi: 10.1016/j.compag.2021.106683.

[21] Y. Egi, M. Hajyzadeh, and E. Eyceyurt, "Drone-Computer Communication Based Tomato Generative Organ Counting Model Using YOLO V5 and Deep-Sort," *Agriculture (Switzerland)*, vol. 12, no. 9, Sep. 2022, doi: 10.3390/agriculture12091290.

[22] E. Bayraktar, M. E. Basarkan, and N. Celebi, "A low-cost UAV framework towards ornamental plant detection and counting in the wild," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 167, pp. 1–11, Sep. 2020, doi: 10.1016/j.isprsjprs.2020.06.012.

[23] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," 2004.

[24] P. Fernández Alcantarilla, A. Bartoli, and A. J. Davison, "LNCS 7577 - KAZE Features," 2012.

[25] C. Cortes, V. Vapnik, and L. Saitta, "Support-Vector Networks Editor," Kluwer Academic Publishers, 1995.

[26] "GitHub - wkentaro/labelme: Image Polygonal Annotation with Python (polygon, rectangle, circle, line, point and image-level flag annotation)." https://github.com/wkentaro/labelme (accessed Jan. 23, 2022).

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Cham: Springer International Publishing, 2015, pp. 234–241.

[28] E. ik Jeon, S. Kim, S. Park, J. Kwak, and I. Choi, "Semantic segmentation of seagrass habitat from drone imagery based on deep learning: A comparative

study," *Ecol Inform*, vol. 66, no. August, p. 101430, 2021, doi: 10.1016/j.ecoinf.2021.101430.

[29] S. I. Moazzam, U. S. Khan, W. S. Qureshi, T. Nawaz, and F. Kunwar, "Towards automated weed detection through two-stage semantic segmentation of tobacco and weed pixels in aerial Imagery," *Smart Agricultural Technology*, p. 100142, 2022.

[30] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.03385

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 2015, [Online]. Available: http://arxiv.org/abs/1512.00567

[33] "Welcome To Colaboratory - Colaboratory." https://colab.research.google.com/?utm_source=scs-index (accessed Jan. 23, 2022).

[34] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors."

[35] "GitHub - heartexlabs/labelImg: LabelImg is now part of the Label Studio community. The popular image annotation tool created by Tzutalin is no longer actively being developed, but you can check out Label Studio, the open source data labeling tool for images, text, hypertext, audio, video and time-series data." https://github.com/heartexlabs/labelImg (accessed Jan. 07, 2023).

[36] "Agisoft Metashape: Agisoft Metashape." https://www.agisoft.com/ (accessed Jun. 29, 2023).